Reinhard Mennicken
Ekkehard Wagenführer

Numerische Mathematik 2

Mathematik
Grundkurs

ro
ro
ro
vieweg

Dr. rer. nat. Reinhard Mennicken ist Professor
im Fachbereich Mathematik der Universität Regensburg

Dr. rer. nat. Ekkehard Wagenführer ist Akademischer Rat
im Fachbereich Mathematik der Universität Regensburg

(Kurzbiographien der Autoren stehen auf Seite 252)

Redaktion: Verlag Vieweg, Wiesbaden

Veröffentlicht im Rowohlt Taschenbuch Verlag GmbH,
Reinbek bei Hamburg, November 1977

Umschlagentwurf Werner Rebhuhn
Satz Viewig, Braunschweig
Druck Clausen & Bosse, Leck/Schleswig
Printed in Germany
1480-ISBN 3 499 27034 x

Inhaltsverzeichnis

Vorwort

Der vorliegende Band „Numerische Mathematik 2" behandelt die Themen Eigenwertberechnung, Iterationsverfahren und Interpolation. Wie auch bereits im 1. Band, legen die Autoren großes Gewicht auf die Bereitstellung theoretischer Grundlagen; diese sind teilweise auch über die Numerik hinaus für andere Gebiete der Angewandten Mathematik von Interesse, so beispielsweise der Banachsche Fixpunktsatz oder der Begriff der Differenzierbarkeit in normierten Vektorräumen.

Die numerischen Verfahren sind wiederum ausführlich hinsichtlich ihrer Anwendbarkeit diskutiert, ferner sind zur Erläuterung zahlreiche Beispiele angegeben. Die Algorithmen sind so formuliert, daß man hiernach ohne Schwierigkeiten Computerprogramme schreiben kann.

Auf Grund der fortlaufenden Numerierung der Kapitel sind Zitate aus Band 1, d.h. aus den Kapiteln 1–4, auch ohne entsprechenden Hinweis an den angegebenen Nummern zu erkennen.

Danken möchten die Autoren allen, die an diesem Band mitgewirkt haben, insbesondere unseren Mitarbeitern Dr. *B. Sagraloff* und Dipl.-Math. *J. Wiesmüller*, die einen Teil der Korrekturen übernommen haben. Ferner danken wir unseren Diplomanden, die die einzelnen Algorithmen programmiert und uns damit die Zahlenwerte zu den numerischen Beispielen geliefert haben. Schließlich gilt unser Dank dem Vieweg Verlag für die gute Zusammenarbeit.

Regensburg, im September 1977

R. Mennicken
E. Wagenführer

5. Eigenwertberechnung bei Matrizen

Es sei $A \in M(n \times n, \mathbb{C})$. Ein $\lambda \in \mathbb{C}$ heißt *Eigenwert* der Matrix A, wenn ein $x \in \mathbb{C}^n, \neq 0$ mit der Eigenschaft

$$Ax = \lambda x$$

existiert; weiter nennt man jedes derartige x einen *Eigenvektor* zum Eigenwert λ.

Ziel dieses Kapitels ist es, die wichtigsten numerischen Verfahren zur Berechnung der Eigenwerte und Eigenvektoren mitzuteilen. Wegen der Verschiedenartigkeit der einzelnen Methoden, die eine Verzahnung untereinander nicht ausschließt, beschreiben wir hier aus Gründen der Übersichtlichkeit kurz die Inhalte der folgenden Abschnitte.

In 5.1 ist die Potenzmethode dargestellt. Sie ist unter gewissen Voraussetzungen dazu geeignet, eine Näherung für den betraglich größten Eigenwert und einen zugehörigen Eigenvektor zu bestimmen; dabei wird auch auf die inverse Potenzmethode eingegangen. Zum tieferen Verständnis stellen wir diesen Verfahren einige grundlegende Aussagen über Systeme linearer Differenzengleichungen erster Ordnung voran.

Als Anwendung behandeln wir in 5.2 das Bernoulli-Verfahren. Es dient dazu, die betraglich größte Nullstelle eines Polynoms näherungsweise zu berechnen. Zusätzlich notieren wir – ausgehend von den Ergebnissen in 5.1 – einige wichtige Sätze über lineare Differenzengleichungen n-ter Ordnung, die über den Rahmen dieses Kapitels hinaus, nämlich beispielsweise bei der numerischen Behandlung von Differentialgleichungen (in Band 3) von Bedeutung sind.

Bekanntlich ist – vgl. z.B. Fischer [17], S. 166 – λ genau dann Eigenwert von A, wenn λ eine Nullstelle des charakteristischen Polynoms

$$\varphi(\mu) = \det(\mu I - A)$$

ist. Damit ist die Eigenwertberechnung auf die Nullstellenbestimmung eines Polynoms zurückgeführt. Allerdings ist es schon im Falle $n = 4$ ein mühsames Unterfangen, die Koeffizienten von $\varphi(\mu)$ zu berechnen. Dieser Schwierigkeit begegnet man, indem man von A zu einer ähnlichen Matrix $B = T^{-1}AT$ möglichst einfacher Form übergeht und dabei die Beziehungen

$$\begin{cases} \det(\mu I - B) = \det(\mu I - A), \\ \lambda \text{ Eigenwert von } B \Leftrightarrow \lambda \text{ Eigenwert von } A, \\ x \text{ Eigenvektor von } B \Leftrightarrow Tx \text{ Eigenvektor von } A \end{cases}$$

beachtet.

Im Hinblick darauf geben wir in 5.3 finite Verfahren an, die A durch Ähnlichkeitstransformationen T auf eine obere Hessenberg- oder sogar eine Tridiagonalmatrix B bringen. Wie weiter in 5.4 gezeigt wird, erlauben Matrizen dieser Art die Berechnung des zugehörigen charakteristischen Polynoms ohne größeren Aufwand. Es bleibt natürlich dann das Problem, die Nullstellen des charakteristischen Polynoms zu bestimmen.

In 5.5 bringen wir das QR- und das LR-Verfahren, anschließend in 5.6 das Jacobi-Verfahren. Hierbei handelt es sich um infinite Methoden: mittels einer Folge ähnlicher Transformationen wird A in eine Dreiecks- bzw. Diagonalmatrix übergeführt. Diese Verfahren werden insbesondere in dem Fall herangezogen, daß sämtliche Eigenwerte einer Matrix berechnet werden sollen.

In 5.7 schließlich beschäftigen wir uns mit Fehlerbetrachtungen.

5.1. Lineare Differenzengleichungen erster Ordnung im $\mathbb{C}^n$. Potenzmethode

Es sei $(A_k)_0^\infty$ eine Folge komplexer (n,n)-Matrizen, $(b_k)_0^\infty$ eine Folge im $\mathbb{C}^n$. Dann heißt ein System von Gleichungen der Form

$$(5.1.1) \qquad y_{k+1} - A_k y_k = b_k \qquad (k = 0, 1, 2, \ldots)$$

eine *lineare Differenzengleichung erster Ordnung im* $\mathbb{C}^n$; diese nennt man homogen, falls sämtliche b_k verschwinden, ansonsten inhomogen.

Wie üblich bezeichnen wir mit $(\mathbb{C}^n)^{\mathbb{N}}$ die Menge aller Folgen $(y_k)_0^\infty$ im $\mathbb{C}^n$. Diese Menge wird bekanntlich zu einem Vektorraum über $\mathbb{C}$, wenn man die Addition und die Multiplikation mit $\alpha \in \mathbb{C}$ komponentenweise erklärt.

Wir notieren einige elementare Eigenschaften von (5.1.1) in

(5.1.2) **Hilfssatz**

(i) *Zu jedem „Anfangswert"* $y_0 \in \mathbb{C}^n$ *existieren eindeutig* $y_1, y_2, \ldots, \in \mathbb{C}^n$, *so daß* (5.1.1) *erfüllt ist.*

(ii) *Es ist*

$$\Delta (y_k)_0^\infty := (y_{k+1} - A_k y_k)_0^\infty$$

eine lineare Abbildung von $(\mathbb{C}^n)^{\mathbb{N}}$ *in* $(\mathbb{C}^n)^{\mathbb{N}}$.

(iii) *Es bezeichne*

$$\Lambda := \{(y_k)_0^\infty : y_{k+1} - A_k y_k = 0 \quad (k = 0, 1, 2, \ldots)\},$$

also den Nullraum der Abbildung Δ; Λ *ist ein Unterraum des* $(\mathbb{C}^n)^{\mathbb{N}}$ *mit*

$$\dim \Lambda = n .$$

(iv) *Ist* $(y_k)_0^\infty$ *eine Lösung von* (5.1.1), *so genügt eine Folge* $(z_k)_0^\infty$ *genau dann* (5.1.1), *wenn*

$$(y_k - z_k)_0^\infty = (y_k)_0^\infty - (z_k)_0^\infty \in \Lambda .$$

Beweis. Die Aussage (i) ist trivial, da man jede der Gleichungen (5.1.1) nach y_{k+1} auflösen kann. Zu (ii) notieren wir mit $\alpha, \beta \in \mathbb{C}$ sowie $(y_k)_0^\infty, (z_k)_0^\infty \in (\mathbb{C}^n)^{\mathbb{N}}$

$$\Delta(\alpha(y_k)_0^\infty + \beta(z_k)_0^\infty) = \Delta(\alpha y_k + \beta z_k)_0^\infty = (\alpha y_{k+1} + \beta z_{k+1} - A_k(\alpha y_k + \beta z_k))_0^\infty$$
$$= \alpha(y_{k+1} - A_k y_k)_0^\infty + \beta(z_{k+1} - A_k z_k)_0^\infty = \alpha\Delta(y_k)_0^\infty + \beta\Delta(z_k)_0^\infty .$$

Bezüglich (iii) stellen wir fest, daß Λ als Nullraum einer linearen Abbildung ein Unterraum ist; weiter betrachten wir die Abbildung

$$\varphi: \Lambda \to \mathbb{C}^n, \qquad (y_k)_0^\infty \mapsto y_0 .$$

Offensichtlich ist φ linear; ferner ist φ nach (i), angewendet auf $b_k = 0$ $(k = 0, 1, 2, \ldots)$ bijektiv, mithin ein Vektorraum-Isomorphismus. Daher gilt, wie behauptet,

$$\dim \Lambda = \dim \mathbb{C}^n = n .$$

Schließlich ist (iv) nach (ii) klar.

(5.1.3) **Definition.** Jede Basis von Λ heißt *Fundamentalsystem* der zu (5.1.1) zugehörigen homogenen Differenzengleichung.

Dem Beweis von (5.1.2), (iii) entnehmen wir die

(5.1.4) **Bemerkung.** Ist $(a_1, a_2, \ldots, a_n)$ eine beliebige Basis des $\mathbb{C}^n$, so bilden die n Folgen $(y_k^{(\nu)})_{k=0}^\infty$ $(\nu = 1, 2, \ldots, n)$, definiert durch

$$y_0^{(\nu)} = a_\nu ,$$
$$y_{k+1}^{(\nu)} = A_k y_k^{(\nu)} \qquad (k = 0, 1, 2, \ldots) ,$$

ein Fundamentalsystem der zu (5.1.1) zugehörigen homogenen Differenzengleichung.

Zum *Beweis* beachtet man, daß

$$(y_k^{(\nu)})_{k=0}^\infty = \varphi^{-1}(a_\nu) \qquad (\nu = 1, 2, \ldots, n) .$$

Man nennt (5.1.1) eine Differenzengleichung *mit konstanten Koeffizienten*, wenn die A_k von k unabhängig sind, d.h. wenn mit einer festen (n,n)-Matrix A für $k = 0, 1, 2, \ldots$ $A_k = A$ gilt; dabei wird nicht ausgeschlossen, daß die b_k von k abhängen.

Im folgenden beschäftigen uns homogene lineare Differenzengleichungen mit konstanten Koeffizienten, d.h. wir betrachten bei vorgegebenem $A \in M(n \times n, \mathbb{C})$ und $y_0 \in \mathbb{C}^n$ die Rekursionen

(5.1.5) $\qquad y_{k+1} = Ay_k \qquad (k = 0, 1, 2, \ldots) .$

Als Lösung erhält man trivialerweise

(5.1.6) $\qquad y_k = A^k y_0 \qquad (k = 0, 1, 2, \ldots) .$

Die Berechnung dieser Vektoren y_k (für geeignete $y_0 \in \mathbb{C}^n$) heißt *Potenzmethode* oder *von-Mises-Verfahren;* man nennt sie manchmal auch – vgl. 5.2 – *verallgemeinertes Bernoulli-Verfahren.* Zur Herleitung einer Konvergenzaussage benötigen wir eine explizitere Darstellung der Lösung als (5.1.6). Hierzu zitieren wir aus der Linearen Algebra – siehe [21], S. 208 – den

(5.1.7) **Satz** *(über die Jordansche Normalform). Es seien* $\lambda_1, \ldots, \lambda_m$ *die verschiedenen Eigenwerte der* (n,n)*-Matrix* A. *Für* $\mu \in \{1, 2, \ldots, m\}$ *bezeichne*

$$H_\mu := \{h \in \mathbb{C}^n: \exists k \in \mathbb{N}\ (A - \lambda_\mu I)^k h = 0\}$$

den Hauptraum sowie

$$E_\mu := \{x \in \mathbb{C}^n: (A - \lambda_\mu I)\, x = 0\}$$

den Eigenraum zum Eigenwert λ_μ.

Wir stellen fest:

(i) *Es gilt die direkte Zerlegung*

$$\mathbb{C}^n = H_1 \dotplus H_2 \dotplus \ldots \dotplus H_m .$$

(ii) *Jeder Hauptraum* H_μ *besitzt eine weitergehende Zerlegung*

$$H_\mu = H_{\mu,1} \dotplus H_{\mu,2} \dotplus \ldots \dotplus H_{\mu,k_\mu} ,$$

wobei jeder dieser Unterräume $H_{\mu,i}$ *durch eine Basis* $(u_1, \ldots, u_{p_{\mu,i}})$ *aus „aufsteigenden" Hauptvektoren, d.h. mit*

$$(A - \lambda_\mu I)\, u_1 = 0; \quad (A - \lambda_\mu I)\, u_{j+1} = u_j \qquad (j = 1, 2, \ldots, p_{\mu,i} - 1)$$

charakterisiert ist. (Die Abhängigkeit der u_j *von* μ *und* i *ist dabei nicht markiert.)*

(iii) E_μ *ist ein Unterraum von* H_μ *mit*

$$\dim E_\mu = k_\mu \geq 1 .$$

Ist $h \in H_\mu$, so nennt man

(5.1.8) $$p := \min \{k \in \mathbb{N}: (A - \lambda_\mu I)^k h = 0\}$$

die *Stufe* des Hauptvektors h. Offensichtlich gilt $p = 0$ genau dann, wenn $h = 0$ ist. Hauptvektoren der Stufe $p \geq 1$ sind durch die Eigenschaften

$$(A - \lambda_\mu I)^{p-1} h \neq 0, \quad (A - \lambda_\mu I)^p h = 0$$

gekennzeichnet; die Hauptvektoren der Stufe 1 sind mithin gerade die Eigenvektoren. Da jedes u aus der in (5.1.7), (ii) angegebenen Basis von H_μ mit

$$q_\mu = \max_{i=1}^{k_\mu} p_{\mu,i} \leq \dim H_\mu \leq n$$

die Gleichung $(A - \lambda_\mu I)^{q_\mu} h = 0$ erfüllt, ergibt sich gemäß Definition (5.1.8) unmittelbar die Beziehung $p \leq n$.

(5.1.9) **Hilfssatz.** *Es sei* $h \in H_\mu$ *ein Hauptvektor der Stufe* p *und hierzu*

$$h^{(0)} := h, \quad h^{(\nu+1)} := (A - \lambda_\mu I)\, h^{(\nu)} = (A - \lambda_\mu I)^{\nu+1} h \qquad (\nu = 0, 1, 2, \ldots).$$

Dann gilt für $k = 0, 1, 2, \ldots$

$$A^k h = \sum_{\nu=0}^{k} \binom{k}{\nu} \lambda_\mu^{k-\nu} h^{(\nu)} = \sum_{\nu=0}^{\min\{k, p-1\}} \binom{k}{\nu} \lambda_\mu^{k-\nu} h^{(\nu)}.$$

Den *Beweis* führen wir mit Induktion über k. Für $k = 0$ haben wir

$$A^0 h = h = h^{(0)};$$

der Schluß von k auf $k + 1$ ist mit

$$A^{k+1} h = A(A^k h) = \sum_{\nu=0}^{k} \binom{k}{\nu} \lambda_\mu^{k-\nu} \{(A - \lambda_\mu I)\, h^{(\nu)} + \lambda_\mu h^{(\nu)}\}$$

$$= \sum_{\nu=0}^{k} \binom{k}{\nu} \lambda_\mu^{k-\nu} h^{(\nu+1)} + \sum_{\nu=0}^{k} \binom{k}{\nu} \lambda_\mu^{(k+1)-\nu} h^{(\nu)}$$

$$= \sum_{\nu=1}^{k+1} \binom{k}{\nu-1} \lambda_\mu^{(k+1)-\nu} h^{(\nu)} + \sum_{\nu=0}^{k} \binom{k}{\nu} \lambda_\mu^{(k+1)-\nu} h^{(\nu)}$$

$$= \sum_{\nu=0}^{k+1} \binom{k+1}{\nu} \lambda_\mu^{(k+1)-\nu} h^{(\nu)}$$

erbracht.

Nach diesen Vorbereitungen beweisen wir über die Konvergenz des von-Mises-Verfahrens den folgenden

(5.1.10) **Satz.** *Es sei* $A \in M(n \times n, \mathbb{C})$. $\lambda_1, \lambda_2, \ldots, \lambda_m$ *seien die verschiedenen Eigenwerte von* A, *und zwar mit der Eigenschaft*

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_m|;$$

insbesondere sei mithin $m \geq 2$. *Wir setzen*

$$m' := \begin{cases} m, & \textit{falls} \quad \lambda_m \neq 0, \\ m-1 & \textit{sonst} \end{cases}$$

sowie

$$\rho := \left| \frac{\lambda_2}{\lambda_1} \right|.$$

Der „Startvektor" $y_0 \in \mathbb{C}^n$ *besitze mit* $h_1 \in E_1, \neq 0$, $h_\mu \in H_\mu$ $(\mu = 2, \ldots, m)$ *die (eindeutige) Zerlegung*

$$y_0 = h_1 + h_2 + \ldots + h_m.$$

Schließlich bezeichne p_μ *die Stufe des Hauptvektors* h_μ *und*

$$p := \max \{p_\mu : 1 \leqq \mu \leqq m'\}.$$

Wir behaupten:

(i) *Die* y_k, *definiert durch* (5.1.5) *bzw.* (5.1.6) *genügen einer Darstellung*

$$y_k = \lambda_1^k (h_1 + v_k),$$

wobei die v_k *mit einem* $\gamma \geqq 0$ *für* $k = 1, 2, \ldots$ *in der Form*

$$\|v_k\|_\infty \leqq \gamma \cdot \rho^k k^{p-1}$$

abschätzbar sind, also insbesondere gegen Null konvergieren.

(ii) *Setzt man* $h_1 = (\eta_1^{(j)})_{j=1}^n$, $y_k = (\xi_k^{(j)})_{j=1}^n$, *so gilt für jedes* $j \in \{1, \ldots, n\}$ *mit* $\eta_1^{(j)} \neq 0$

$$\lim_{k \to \infty} \frac{\xi_{k+1}^{(j)}}{\xi_k^{(j)}} = \lambda_1$$

bzw. genauer

$$\frac{\xi_{k+1}^{(j)}}{\xi_k^{(j)}} = \lambda_1 + O(\rho^k k^{p-1}) \qquad (k \to \infty).$$

(iii) *Definiert man für* $k \in \mathbb{N}$

$$j_k := \min \{i \in \{1, \ldots, n\} : |\xi_k^{(i)}| = \|y_k\|_\infty\},$$

so hat man $\xi_k^{(j_k)} \neq 0$ *für* $k \in \mathbb{N}$ *und*

$$\lim_{k \to \infty} \frac{\xi_{k+1}^{(j_k)}}{\xi_k^{(j_k)}} = \lambda_1$$

bzw. wieder genauer

$$\frac{\xi_{k+1}^{(j_k)}}{\xi_k^{(j_k)}} = \lambda_1 + O(\rho^k k^{p-1}) \qquad (k \to \infty).$$

Beweis. Nach (5.1.6), (5.1.9) gilt für $k \geqq \max_{\mu=1}^{m} p_\mu$, also sicher für $k \geqq n$

$$y_k = A^k y_0 = \sum_{\mu=1}^{m} \left(\sum_{\nu=0}^{p_\mu - 1} \binom{k}{\nu} \lambda_\mu^{k-\nu} h_\mu^{(\nu)} \right) \tag{5.1.11}$$

und daher für diese k auf Grund der Definition von m' und wegen $p_1 = 1$

$$y_k = A^k y_0 = \lambda_1^k (h_1 + v_k) \tag{5.1.12}$$

mit

$$(5.1.13)\qquad v_k = \sum_{\mu=2}^{m'} \left(\frac{\lambda_\mu}{\lambda_1}\right)^k \sum_{\nu=0}^{p_\mu - 1} \binom{k}{\nu} \lambda_\mu^{-\nu} h_\mu^{(\nu)} .$$

Unter Beachtung der Ungleichung $\binom{k}{\nu} \leqq k^\nu$ schätzt man

$$(5.1.14)\qquad \|v_k\|_\infty \leqq \rho^k k^{p-1} \sum_{\mu=2}^{m'} \sum_{\nu=0}^{p_\mu - 1} |\lambda_\mu|^{-\nu} \|h_\mu^{(\nu)}\|_\infty = \gamma \rho^k k^{p-1}$$

ab, womit (i) bereits bewiesen ist.

Zum Beweis von (ii) setzen wir $v_k = (\epsilon_k^{(j)})_{j=1}^n$ und folgern aus (i) für $j = 1, \ldots, n,\ k \geqq n$

$$\xi_k^{(j)} = \lambda_1^k (\eta_1^{(j)} + \epsilon_k^{(j)})$$

mit

$$\left|\epsilon_k^{(j)}\right| \leqq \|v_k\|_\infty \to 0 \qquad (k \to \infty) .$$

Zu

$$\eta := \min \{|\eta_1^{(j)}| : j \in \{1, \ldots, n\},\ \eta_1^{(j)} \neq 0\}$$

wählen wir ein $k_0 \geqq n$, so daß für $k \geqq k_0$, $j = 1, \ldots, n$

$$|\epsilon_k^{(j)}| \leqq \|v_k\|_\infty < \tfrac{1}{2}\eta ,$$

mithin

$$(5.1.15)\qquad |\eta_1^{(j)} + \epsilon_k^{(j)}| \begin{cases} \geqq \eta - |\epsilon_k^{(j)}| > \tfrac{1}{2}\eta , & \text{falls } \eta_1^{(j)} \neq 0 , \\ = |\epsilon_k^{(j)}| < \tfrac{1}{2}\eta & \text{sonst} \end{cases}$$

erfüllt ist. Ist nun $\eta_1^{(j)} \neq 0$, so können wir für $k \geqq k_0$ durch $\xi_k^{(j)}$ dividieren und erhalten

$$\frac{\xi_{k+1}^{(j)}}{\xi_k^{(j)}} = \lambda_1 \frac{\eta_1^{(j)} + \epsilon_{k+1}^{(j)}}{\eta_1^{(j)} + \epsilon_k^{(j)}} \to \lambda_1 \qquad (k \to \infty)$$

bzw. nach (5.1.14) mit einem geeigneten $\gamma' \geqq 0$ genauer

$$\left| \frac{\xi_{k+1}^{(j)}}{\xi_k^{(j)}} - \lambda_1 \right| = |\lambda_1| \frac{|\epsilon_k^{(j)} - \epsilon_{k+1}^{(j)}|}{|\eta_1^{(j)} + \epsilon_k^{(j)}|} \leqq |\lambda_1| \frac{\|v_k\|_\infty + \|v_{k+1}\|_\infty}{\frac{1}{2}\eta} \leqq \gamma' \rho^k k^{p-1} .$$

Zu (iii) beachten wir, daß gemäß (5.1.13) $v_k \in (H_2 \dot{+} \ldots \dot{+} H_m)$ und nach Voraussetzung $h_1 \in H_1$, $\neq 0$ gilt; wegen $H_1 \cap (H_2 \dot{+} \ldots \dot{+} H_m) = \{0\}$ ist infolgedessen für $k \in \mathbb{N}$ $y_k \neq 0$ und damit $\xi_k^{(j_k)} \neq 0$. Den Ungleichungen $|\xi_k^{(j_k)}| \geqq |\xi_k^{(j)}|$

$(j = 1, \ldots, n)$ entnehmen wir weiter, daß für alle $k \geqq k_0$ $\eta_1^{(j_k)} \neq 0$ ist; sonst hätte man nämlich für ein $j \in \{1, \ldots, n\}$ mit $\eta_1^{(j)} \neq 0$ nach (5.1.15) den Widerspruch

$$\tfrac{1}{2}\eta > |\epsilon_k^{(j_k)}| = |\eta_1^{(j_k)} + \epsilon_k^{(j_k)}| \geqq |\eta_1^{(j)} + \epsilon_k^{(j)}| \geqq \eta - |\epsilon_k^{(j)}| > \tfrac{1}{2}\eta\,.$$

Es folgt, daß die in (ii) gewonnenen Abschätzungen bezüglich $j = j_k$ gelten, womit der Satz (5.1.10) vollständig bewiesen ist.

Wir bemerken, daß die angegebenen Abschätzungen kaum für eine Fehlerrechnung zu verwenden sind, da die dort auftretenden Konstanten nicht a priori bekannt sind. Die Abschätzungen beinhalten jedoch eine Aussage über die Güte der Konvergenz: eine brauchbare Näherung von λ_1 wird man umso eher erreichen, je kleiner ρ ist.

Die Voraussetzung $m \geqq 2$ wird in (5.1.10) ebenso wie an einigen Stellen im Folgenden ausschließlich zur Ausschaltung eines trivialen Sonderfalls getroffen. Es ist nämlich im Fall $m = 1$ $0 \neq y_0 \in E_1$, $v_k = 0$ $(k \in \mathbb{N})$ sowie für $j \in \{1, \ldots, n\}$ mit $\eta_1^{(j)} \neq 0$

$$\frac{\xi_{k+1}^{(j)}}{\xi_k^{(j)}} = \lambda_1 \qquad (k \in \mathbb{N})\,.$$

Die Bedingung $h_1 \in E_1$ ist im Fall $H_1 = E_1$ immer erfüllt; es fragt sich nur, wie man dann zu einem y_0 mit $h_1 \neq 0$ kommt. Hätte man $h_1 = 0$ in der Darstellung von y_0 als Summe von Hauptvektoren, gleichzeitig etwa $0 \neq h_2 \in E_2$, $|\lambda_2| > |\lambda_3|$, so würde das Verfahren rein theoretisch den Eigenwert λ_2 statt λ_1 liefern. Während die theoretischen y_k sämtlich in $H_2 \dotplus \ldots \dotplus H_m$ liegen, erhält man durch Rundungsfehler schon nach 1 bis 2 Iterationen ein numerisch berechnetes

$$\tilde{y}_l \in H_1 \dotplus H_2 \dotplus \ldots \dotplus H_m$$

mit einem nicht verschwindenden Anteil in H_1. Gemäß (5.1.12)–(5.1.14), auf $\tilde{y}_l$ statt y_0 angewendet, wächst durch den Faktor λ_1^k die ursprünglich kleine Komponente in H_1 stärker als die Anteile in $H_2 \dotplus \ldots \dotplus H_m$. Man kann also $y_0 \in \mathbb{C}^n$, $\neq 0$ beliebig wählen; ein Anteil in H_1 entsteht nach einigen Iterationen.

Da die Komponenten von y_k wie λ_1^k wachsen bzw. fallen und infolgedessen die Gefahr des Exponentenüberlaufs besteht, berechnet man bei der praktischen Anwendung an Stelle der y_k normierte Vektoren $\hat{y}_k = \gamma_k y_k$. Wir notieren dazu den

(5.1.16) *Algorithmus.* Wir setzen $y_0' := y_0$ und konstruieren für $k = 0, 1, 2, \ldots$ rekursiv: Ist $y_k' = (\xi_k^{(i)'})_{i=1}^n$ gegeben, so bestimmen wir

$$i_k := \min\{i \in \{1, \ldots, n\} : |\xi_k^{(i)'}| = \|y_k'\|_\infty\}$$

und berechnen

$$\hat{y}_k := \frac{1}{\xi_k^{(i_k)'}}\, y_k'\,, \qquad y_{k+1}' := A\hat{y}_k\,.$$

Schematisch sieht das folgendermaßen aus:

$$y_0 = y_0' \to \hat{y}_0 \to A\hat{y}_0 = y_1' \to \hat{y}_1 \to A\hat{y}_1 = y_2' \to \hat{y}_2 \quad \text{usf.}$$

Den Zusammenhang mit den y_k stellen wir her im

(5.1.17) **Hilfssatz.** *Es gilt für alle* $k \in \mathbb{N}$ *mit gewissen* $\alpha_k \in \mathbb{C}, \neq 0$ $y_k' = \alpha_k y_k$, *ferner im Sinne von* (5.1.10), (iii) $i_k = j_k$ *sowie schließlich*

$$\hat{y}_k = \frac{1}{\xi_k^{(j_k)}} y_k .$$

Den *Beweis* führen wir mit Induktion über k. Der Fall $k = 0$ ist wegen $y_0' = y_0$ klar. Beim Schluß von k auf $k + 1$ erhält man mit einem $\gamma_k \in \mathbb{C}, \neq 0$ $\hat{y}_k = \gamma_k y_k$ und infolgedessen

$$y_{k+1}' = A\hat{y}_k = \gamma_k A y_k = \gamma_k y_{k+1} .$$

Hieraus folgt der Reihe nach

$$i_{k+1} = j_{k+1}, \qquad \xi_{k+1}^{(j_{k+1})'} = \gamma_k \xi_{k+1}^{(j_{k+1})}, \qquad \hat{y}_{k+1} = \frac{1}{\xi_{k+1}^{(j_{k+1})}} y_{k+1} .$$

Zur Konvergenz des Algorithmus (5.1.16) vermerken wir den

(5.1.18) **Satz.** *Unter den Voraussetzungen des Satzes* (5.1.10) *gilt*

(i) $$\xi_{k+1}^{(j_k)'} = \frac{\xi_{k+1}^{(j_k)}}{\xi_k^{(j_k)}} \to \lambda_1 \qquad (k \to \infty)$$

sowie mit geeigneten $\beta_k \in \mathbb{C}, \neq 0$

(ii) $$\|\hat{y}_k - \beta_k h_1\|_\infty = O(\rho^k k^{p-1}) ,$$

mithin insbesondere $\hat{y}_k - \beta_k h_1 \to 0$ *für* $k \to \infty$.

Beweis. Die Aussage (i) folgt unmittelbar aus

$$y_{k+1}' = A\hat{y}_k = \frac{1}{\xi_k^{(j_k)}} y_{k+1} \qquad (k \in \mathbb{N}) .$$

Zu (ii) notieren wir – vgl. den Beweis des Satzes (5.1.10) – die Beziehung

$$\hat{y}_k = \frac{1}{\xi_k^{(j_k)}} y_k = \frac{\lambda_1^k}{\lambda_1^k (\eta_1^{(j_k)} + \epsilon_k^{(j_k)})} (h_1 + v_k) =: \beta_k (h_1 + v_k) .$$

Auf Grund von (5.1.15) können wir für $k \geq k_0$ $|\beta_k| \leqq \frac{2}{\eta}$ abschätzen und erhalten daher mit (5.1.14) für diese k

$$\|\hat{y}_k - \beta_k h_1\|_\infty = \|\beta_k v_k\|_\infty \leqq \frac{2}{\eta} \gamma \rho^k k^{p-1} \qquad (k \geq k_0) .$$

Der Abstand der normierten Vektoren $\hat{y}_k$ zum Eigenraum E_1 konvergiert also gegen Null. Daß in den meisten Fällen die $\hat{y}_k$ sogar gegen einen festen Eigenvektor konvergieren, zeigen wir in dem

(5.1.19) **Zusatz.** Falls h_1 nur eine betraglich maximale Komponente besitzt, also ein $i_0 \in \{1, \ldots, n\}$ mit

$$|\eta_1^{(i_0)}| > |\eta_1^{(i)}| \qquad (i \neq i_0)$$

existiert, so gilt ab einem $k_0 \in \mathbb{N}$

$$j_k = i_0 \qquad (k \geq k_0)$$

und folglich

$$\lim_{k \to \infty} \hat{y}_k = \frac{1}{\eta_1^{(i_0)}} h_1 .$$

Beweis. Wir setzen

$$\delta := \min \{ |\eta_1^{(i_0)}| - |\eta_1^{(i)}| : i \neq i_0 \}$$

und bestimmen hierzu $k_0 \in \mathbb{N}$ mit der Eigenschaft

$$\|v_k\|_\infty < \frac{\delta}{2} \qquad (k \geq k_0) .$$

Dann gilt für $i \neq i_0$, $k \geq k_0$

$$\left|\xi_k^{(i)}\right| = \left|\lambda_1^k\right| \left|\eta_1^{(i)} + \epsilon_k^{(i)}\right| \leq \left|\lambda_1^k\right| \left(\left|\eta_1^{(i)}\right| + \|v_k\|_\infty\right) < \left|\lambda_1^k\right| \left(\left|\eta_1^{(i)}\right| + \frac{\delta}{2}\right)$$
$$\leq \left|\lambda_1^k\right| \left(\left|\eta_1^{(i_0)}\right| - \frac{\delta}{2}\right) \leq \left|\lambda_1^k\right| \left|\eta_1^{(i_0)} + \epsilon_k^{(i_0)}\right| = \left|\xi_k^{(i_0)}\right|$$

und daher $j_k = i_0$. Wir folgern

$$\hat{y}_k = \frac{1}{\eta_1^{(i_0)} + \epsilon_k^{(i_0)}} (h_1 + v_k) \qquad (k \geq k_0)$$

und daraus unmittelbar die zweite Behauptung.

Im folgenden Beispiel ist die Voraussetzung des Zusatzes erfüllt; gleichzeitig demonstrieren wir, daß auch bei ungünstiger Wahl des Startvektors y_0 das Verfahren den betragsmäßig größten Eigenwert liefert.

(5.1.20) **Zahlenbeispiel.** Vorgegeben sei die symmetrische Matrix

$$A = \begin{pmatrix} 1{,}1600 & 1{,}7280 & 2{,}3040 \\ 1{,}7280 & 1{,}6624 & 0{,}8832 \\ 2{,}3040 & 0{,}8832 & 2{,}1776 \end{pmatrix} .$$

Wir wählen einen Startvektor y_0, der fast orthogonal zu E_1 ist, nämlich

$$y_0^t = (-0{,}73333;\ 1{,}0000;\ -0{,}06667).$$

Wir rechnen doppeltgenau, was einer 16-stelligen Dezimalarithmetik entspricht. Im folgenden sind die Näherungen für λ_1 und die Vektoren $\hat{y}_k$ mit acht Dezimalstellen notiert, wobei die signifikanten Dezimalstellen unterstrichen sind.

k	$\lambda_1^{(k)}$	$\hat{y}_k$		
1	0,33632000	−0,76056496	−0,35343571	1,0
2	0,11310391	−0,79619236	1,0	−0,11103775
3	0,18851106	−0,45982843	−0,15801125	1,0
4	0,97859976	1,0	−0,11623037	0,65346474
5	2,4647367	0,68005302	0,58270594	1,0
6	4,2590881	0,96259512	0,71072068	1,0
7	5,0231277	0,92546636	0,74218031	1,0
8	4,9653681	0,93850615	0,74842515	1,0
9	5,0009273	0,93701539	0,74968509	1,0
10	4,9986053	0,93754025	0,74993700	1,0
11	5,0000371	0,93748061	0,74998740	1,0
12	4,9999441	0,93750161	0,74999748	1,0
13	5,0000015	0,93749922	0,74999950	1,0
14	4,9999978	0,93750006	0,74999990	1,0
15	5,0000001	0,93749997	0,74999998	1,0
16	4,9999999	0,93750000	0,75000000	1,0

Die Näherungswerte zeigen zunächst ein unbestimmtes Verhalten, verursacht durch die Eigenvektoren zu λ_2 und λ_3, bis sich das Verfahren ab etwa der 7. Iteration stabilisiert. Da man schließlich 2–3 Iterationen benötigt, um die Genauigkeit um eine Dezimalstelle zu steigern, empfiehlt es sich, ab etwa der 10. Näherung mit dem Newton-Verfahren weiterzurechnen. Die genauen Werte sind übrigens

$$\lambda_1 = 5{,}0000000,\quad E_1 = \{\alpha \cdot (0{,}93750000;\ 0{,}75000000;\ 1{,}0)^t : \alpha \in \mathbb{C}\}.$$

Zur Konvergenz der Potenzmethode haben wir bisher $|\lambda_1| > |\lambda_2|$, $h_1 \in E_1$ oder schärfer $H_1 = E_1$ vorausgesetzt. In den Fällen $|\lambda_1| = |\lambda_2|$ – oder auch nur $|\lambda_2| \approx |\lambda_1|$ – bzw. $h_1 \in H_1 \setminus E_1$ konvergiert das Verfahren nicht oder zu langsam. Um dennoch λ_1 und einen zugehörigen Eigenvektor zu gewinnen, sind zusätzliche Rechnungen erforderlich; diese Methoden, die einen erheblichen Mehraufwand bedeuten, sind bei Faddejew/Faddejewa [15] eingehend beschrieben.

Wegen der durch $\rho = \frac{|\lambda_2|}{|\lambda_1|}$ bestimmten Konvergenzgeschwindigkeit führt die Potenzmethode nur dann mit vertretbarem Aufwand zum Ziel, wenn $|\lambda_1|$ erheblich

größer als $|\lambda_2|$ ist. Diese Voraussetzung ist für die Iterationsmatrix bei der inversen Potenzmethode von Wielandt [59] im allgemeinen erfüllt. Zunächst notieren wir den

(5.1.21) **Hilfssatz.** *Es sei* $A \in M(n \times n, \mathbb{C})$. $\lambda_1, \ldots, \lambda_m$ *seien die (verschiedenen) Eigenwerte der Matrix* A *und* $H_1, \ldots, H_m$ *die zugehörigen Haupträume. Dann gilt für* $\tilde{\lambda} \in \mathbb{C}, \neq \lambda_\mu$ $(\mu = 1, \ldots, m)$

(i) $A - \tilde{\lambda} I$ *ist invertierbar;*

(ii) $(A - \tilde{\lambda} I)^{-1}$ *besitzt die Eigenwerte* $(\lambda_\mu - \tilde{\lambda})^{-1}$ *und dazu die Haupträume* H_μ; *darüberhinaus ist* $h \in \mathbb{C}^n$ *zur Matrix* A *bezüglich* λ_μ *ein Hauptvektor der Stufe* p *genau dann, wenn* h *zur Matrix* $(A - \tilde{\lambda} I)^{-1}$ *bezüglich* $(\lambda_\mu - \tilde{\lambda})^{-1}$ *ein Hauptvektor der gleichen Stufe ist.*

Beweis. Die Aussage (i) ist klar. Weiter haben wir, da die auftretenden Matrizen vertauschbar sind,

$$(A - \lambda_\mu I)^k = \left\{ (\lambda_\mu - \tilde{\lambda})(A - \tilde{\lambda} I) \left(\frac{1}{\lambda_\mu - \tilde{\lambda}} I - (A - \tilde{\lambda} I)^{-1} \right) \right\}^k$$

$$= (\lambda_\mu - \tilde{\lambda})^k (A - \tilde{\lambda} I)^k \left(\frac{1}{\lambda_\mu - \tilde{\lambda}} I - (A - \tilde{\lambda} I)^{-1} \right)^k$$

und infolgedessen $(A - \lambda_\mu I)^k h = 0$ bzw. $\neq 0$ genau dann, wenn

$$\left(\frac{1}{\lambda_\mu - \tilde{\lambda}} I - (A - \tilde{\lambda} I)^{-1} \right)^k h = 0 \quad \text{bzw.} \quad \neq 0 ,$$

womit auch (ii) bewiesen ist.

Wir folgern den

(5.1.22) **Satz** *(über die inverse Potenzmethode). Die Voraussetzungen des Hilfssatzes* (5.1.21) *seien erfüllt; dabei sei* $m \geq 2$. *Weiter sei* $\tilde{\lambda} \in \mathbb{C}$ *eine Näherung zum Eigenwert* λ_ν *mit*

$$0 < |\tilde{\lambda} - \lambda_\nu| < |\tilde{\lambda} - \lambda_\mu| \qquad (\mu \neq \nu) .$$

Der „Startvektor" $y_0 \in \mathbb{C}^n$ *besitze mit* $h_\nu \in E_\nu, \neq 0$ *sowie* $h_\mu \in H_\mu$ $(\mu \neq \nu)$ *die eindeutige Zerlegung*

$$y_0 = h_1 + h_2 + \ldots + h_m .$$

p_μ *bezeichne die Stufe des Hauptvektors* h_μ *und*

$$p := \max \{ p_\mu : \mu = 1, \ldots, m \}$$

sowie

$$\rho := \max \left\{ \frac{|\tilde{\lambda} - \lambda_\nu|}{|\tilde{\lambda} - \lambda_\mu|} : \mu \neq \nu \right\} .$$

Zu y_0 *definieren wir rekursiv* $(y_k)_0^\infty$, $y_k = (\xi_k^{(j)})_{j=1}^n$ *durch*

$$(A - \tilde{\lambda} I)\, y_{k+1} = y_k$$

und hierzu $(\hat{y}_k)_0^\infty$ *mit den* j_k *aus* (5.1.10), (iii) *durch*

$$\hat{y}_k := \frac{1}{\xi_k^{(j_k)}}\, y_k \,.$$

Dann gilt mit geeigneten $\beta_k \in \mathbb{C}$, $\neq 0$

$$\|\hat{y}_k - \beta_k h_\nu\|_\infty = O(\rho^k k^{p-1}) \,,$$

mithin insbesondere $\hat{y}_k - \beta_k h_\nu \to 0$ *für* $k \to \infty$.

Zum *Beweis* beachten wir, daß die Matrix $(A - \tilde{\lambda} I)^{-1}$, wie man mittels des vorangehenden Hilfssatzes erkennt, den Voraussetzungen des Satzes (5.1.18) genügt.

Die inverse Potenzmethode wird angewendet, wenn man zu einem näherungsweise bekannten Eigenwert – ermittelt z. B. mit den Methoden des Abschnitts 5.4 – einen zugehörigen Eigenvektor sucht. In der Praxis ist beispielsweise $\tilde{\lambda} = \mathrm{rd}(\lambda_\nu)$; es wird dann gegebenenfalls ein Eigenvektor bereits nach wenigen Iterationen erreicht. Da $(A - \tilde{\lambda} I)$ fast singulär ist, ist dafür zu sorgen, daß die Rundungsfehler bei der Lösung der Gleichungssysteme $(A - \tilde{\lambda} I)\, y_{k+1} = y_k$ gering bleiben; daher wird die inverse Potenzmethode meistens auf eine zu A ähnliche Hessenbergmatrix angewendet (vgl. Aufgabe 2.4).

5.2. Lineare Differenzengleichungen n-ter Ordnung in $\mathbb{C}$, Bernoulli-Verfahren

Es seien $(\alpha_{k,0})_{k=0}^\infty, \ldots, (\alpha_{k,n-1})_{k=0}^\infty$ sowie $(\beta_k)_0^\infty$ komplexe Zahlenfolgen. Dann heißt ein System von Gleichungen der Form

$$\eta_{k+n} + \alpha_{k,n-1}\eta_{k+n-1} + \ldots + \alpha_{k,0}\eta_k = \beta_k \qquad (k \in \mathbb{N}) \tag{5.2.1}$$

eine lineare Differenzengleichung n-ter Ordnung in $\mathbb{C}$. Setzt man

$$y_k = \begin{pmatrix} \eta_k \\ \eta_{k+1} \\ \vdots \\ \eta_{k+n-1} \end{pmatrix} \qquad (k \in \mathbb{N}) \,, \tag{5.2.2}$$

so geht (5.2.1) in

$$(5.2.3)\qquad y_{k+1} = \underbrace{\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ & & & & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ -\alpha_{k,0} & -\alpha_{k,1} & \cdots & \cdots & -\alpha_{k,n-1} \end{pmatrix}}_{A_k} y_k + \underbrace{\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \beta_k \end{pmatrix}}_{b_k} \qquad (k \in \mathbb{N}),$$

d.h. in eine lineare Differenzengleichung erster Ordnung im $\mathbb{C}^n$ der Form (5.1.1) über. Diese beiden Differenzengleichungen sind äquivalent im folgenden Sinne:

(5.2.4) **Hilfssatz**

(i) *Ist* $(\eta_k)_0^\infty$ *eine Lösungsfolge von* (5.2.1), *so ist die nach* (5.2.2) *gebildete Folge eine Lösung von* (5.2.3).

(ii) *Ist* $(y_k)_0^\infty$ *eine Lösung von* (5.2.3), *so ist die durch*

$$(5.2.5)\qquad \begin{pmatrix} \eta_0 \\ \vdots \\ \eta_{n-1} \end{pmatrix} := y_0\,, \quad \eta_{n+k} = e_n^t y_{k+1} \qquad (k \in \mathbb{N})$$

definierte Folge $(\eta_k)_0^\infty$ *eine Lösung von* (5.2.1) *und genügt der Beziehung* (5.2.2).

Beweis. Die Aussage (i) weist man durch einfaches Einsetzen nach. Zu (ii) zeigt man die Gültigkeit der Beziehungen (5.2.1), (5.2.2) durch Induktion nach k: (5.2.2) gilt für k = 0 unmittelbar gemäß (5.2.5); (5.2.1) ist für k = 0 mit

$$\eta_n = e_n^t y_1 = e_n^t (A_0 y_0 + b_0) = -\alpha_{0,n-1}\eta_{n-1} - \ldots - \alpha_{0,0}\eta_0 + \beta_0$$

bewiesen. Beim Schluß von k auf k + 1 leitet man zunächst (5.2.2) für k + 1 her; hierzu zieht man die Induktionsvoraussetzung heran und beachtet (5.2.3). Danach ermittelt man (5.2.1) für k + 1 ähnlich wie oben für k = 0.

Auf Grund dieser Einordnung lassen sich die Definitionen und Aussagen des Abschnitts 5.1 sinngemäß auf Differenzengleichungen des Typs (5.2.1) übertragen. Eine genauere Ausführung sei dem Leser überlassen.

Etwas intensiver beschäftigen wir uns mit dem Spezialfall homogener linearer Differenzengleichungen mit konstanten Koeffizienten. Hier ist für $k \in \mathbb{N}$ $\alpha_{k,i} =: a_i$, wonach sich (5.2.1) zu

$$(5.2.6)\qquad \eta_{k+n} + a_{n-1}\eta_{k+n-1} + \ldots + a_0\eta_k = 0 \qquad (k \in \mathbb{N})$$

vereinfacht;

dadurch wird

$$(5.2.7)\qquad A_k = A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ & & & & 0 \\ 0 & \dots & \dots & 0 & 1 \\ -a_0 & \dots & \dots & \dots & -a_{n-1} \end{pmatrix} \qquad (k \in \mathbb{N}),$$

mithin (5.2.3) zu einer Differenzengleichung der speziellen Gestalt (5.1.5).

Zur Ermittlung eines Fundamentalsystems untersuchen wir die Eigenwerte und Haupträume der Matrix A. Hierzu betrachten wir das „charakteristische" Polynom

$$(5.2.8)\qquad P(\lambda) := \lambda^n + a_{n-1}\lambda^{n-1} + \dots + a_0 .$$

$\lambda_1, \dots, \lambda_m$ seien die (verschiedenen) Nullstellen des Polynoms P, q_μ ($\geqq 1$) bezeichne die Ordnung von λ_μ als Nullstelle von P, d.h. es sei $P^{(i)}(\lambda_\mu) = 0$ für $i = 0, 1, \dots, q_\mu - 1$ und $P^{(q_\mu)}(\lambda_\mu) \neq 0$. Für $\mu = 1, \dots, m$ bezeichne

$$(5.2.9)\qquad v_\mu^{(\kappa)} := \left(\binom{k}{\kappa}\lambda_\mu^{k-\kappa}\right)_{k=0}^{n-1} \qquad (\kappa = 0, 1, \dots, q_\mu - 1),$$

wobei die Produkte $\binom{k}{\kappa}\lambda_\mu^{k-\kappa}$ für $k < \kappa$ Null zu setzen sind. Mit diesen Bezeichnungen notieren wir den

(5.2.10) **Hilfssatz.** *Es gilt*

$$P(\lambda) = \det(\lambda I - A).$$

Die Nullstellen λ_μ von P sind mithin die Eigenwerte von A, und zwar der algebraischen Vielfachheit q_μ; zu λ_μ bilden die $v_\mu^{(0)}, \dots, v_\mu^{(q_\mu - 1)}$ eine Basis aufsteigender Hauptvektoren des zugehörigen Hauptraums H_μ.

Beweis. Auf Grund der Produktdarstellung $P(\lambda) = \prod_{\mu=1}^{m} (\lambda - \lambda_\mu)^{q_\mu}$ genügt es, die zweite Aussage nachzuweisen. Hierzu bezeichnen wir bei festem $\mu \in \{1, \dots, m\}$

$$\xi_k^{(\kappa)} := \binom{k}{\kappa}\lambda_\mu^{k-\kappa} \quad (0 \leqq \kappa \leqq q_\mu - 1,\ 0 \leqq k \leqq n-1).$$

Die ersten $n-1$ Komponenten von $(A - \lambda_\mu I)v_\mu^{(0)}$ sind dann gerade

$$\xi_{k+1}^{(0)} - \lambda_\mu \xi_k^{(0)} = \lambda_\mu^{k+1} - \lambda_\mu \lambda_\mu^k = 0 \qquad (k = 0, \dots, n-2);$$

weiter erhalten wir als n-te Komponente

$$-\sum_{k=0}^{n-1} a_k \lambda_\mu^k - \lambda_\mu \lambda_\mu^{n-1} = -P(\lambda_\mu) = 0,$$

also insgesamt

$$(5.2.11) \qquad (A - \lambda_\mu I)\, v_\mu^{(0)} = 0 .$$

Zur Herleitungen der Gleichungen

$$(5.2.12) \qquad (A - \lambda_\mu I)\, v_\mu^{(\kappa)} = v_\mu^{(\kappa-1)} \qquad (\kappa = 1, \ldots, q_\mu - 1)$$

beachten wir, daß in den ersten $n-1$ Zeilen von $(A - \lambda_\mu I)\, v_\mu^{(\kappa)} - v_\mu^{(\kappa-1)}$

$$\xi_{k+1}^{(\kappa)} - \lambda_\mu \xi_k^{(\kappa)} - \xi_k^{(\kappa-1)} = \{\binom{k+1}{\kappa} - \binom{k}{\kappa} - \binom{k}{\kappa-1}\} \lambda_\mu^{k+1-\kappa} = 0 \qquad (k = 0, \ldots, n-2)$$

und in der n-ten Zeile

$$-\sum_{k=\kappa}^{n-1} a_k \xi_k^{(\kappa)} - \lambda_\mu \xi_{n-1}^{(\kappa)} - \xi_{n-1}^{(\kappa-1)} = -\sum_{k=\kappa}^{n-1} \binom{k}{\kappa} a_k \lambda_\mu^{k-\kappa} - \binom{n-1}{\kappa} \lambda_\mu^{n-\kappa} - \binom{n-1}{\kappa-1} \lambda_\mu^{n-\kappa}$$

$$= -\sum_{k=\kappa}^{n} \binom{k}{\kappa} a_k \lambda_\mu^{k-\kappa} = -\frac{1}{\kappa!} P^{(\kappa)}(\lambda_\mu) = 0$$

steht. Gemäß (5.2.11), (5.2.12) bilden die $v_\mu^{(0)}, \ldots, v_\mu^{(q_\mu - 1)}$ eine aufsteigende Folge von Hauptvektoren zum Eigenwert λ_μ und sind als solche linear unabhängig.

Wegen $n = \sum_{\mu=1}^{m} q_\mu$ kann es in den einzelnen Haupträumen H_μ nicht mehr als q_μ linear unabhängige Vektoren geben, so daß die $v_\mu^{(0)}, \ldots, v_\mu^{(q_\mu - 1)}$, wie behauptet, eine Basis von H_μ bilden.

Wir folgern den

(5.2.13) **Satz.** *Die* n *Folgen*

$$(\binom{k}{\kappa} \lambda_\mu^{k-\kappa})_{k=0}^{\infty} \qquad (\kappa = 0, 1, \ldots, q_\mu - 1;\ \mu = 1, \ldots, m)$$

bilden ein Fundamentalsystem von Lösungen zu (5.2.6). *Jede Lösung* $(\eta_k)_0^\infty$ *von* (5.2.6) *besitzt demgemäß mit geeigneten* $\beta_{\mu,\kappa} \in \mathbb{C}$ *die Gestalt*

$$\eta_k = \sum_{\mu=1}^{m} \left(\sum_{\kappa=0}^{q_\mu - 1} \beta_{\mu,\kappa} \binom{k}{\kappa} \lambda_\mu^{k-\kappa} \right) \qquad (k \in \mathbb{N}) .$$

Beweis. Wir betrachten die Lösungsfolge $(\eta_k)_0^\infty$, die bei festen μ, κ gemäß der Bezeichnung (5.2.9) durch die Anfangswerte

$$(\eta_k)_0^{n-1} = v_\mu^{(\kappa)}$$

gegeben ist. Da $v_\mu^{(\kappa)}$ ein Hauptvektor der Stufe $\kappa + 1$ zu λ_μ ist, erhält man mit Hilfssatz (5.1.9) für $k \geq n$

$$A^k v_\mu^{(\kappa)} = \sum_{\nu=0}^{\kappa} \binom{k}{\nu} \lambda_\mu^{k-\nu} (A - \lambda_\mu I)^\nu v_\mu^{(\kappa)} = \sum_{\nu=0}^{\kappa} \binom{k}{\nu} \lambda_\mu^{k-\nu} v_\mu^{(\kappa-\nu)} .$$

Demgemäß ergibt sich unter Berücksichtigung von (5.2.2) für $k = n, n+1, \ldots$

$$\eta_k = e_1^t y_k = e_1^t A^k v_\mu^{(\kappa)} = \sum_{\nu=0}^{\kappa} \binom{k}{\nu} \lambda_\mu^{k-\nu} e_1^t v_\mu^{(\kappa-\nu)} = \binom{k}{\kappa} \lambda_\mu^{k-\kappa} ,$$

und zwar letzteres wegen

$$e_1^t v_\mu^{(\kappa-\nu)} = \begin{cases} 0 & \text{für } 1 \leq \nu \leq \kappa - 1 , \\ 1 & \text{für } \nu = \kappa . \end{cases}$$

Weil die $v_\mu^{(\kappa)}$ eine Basis des $\mathbb{C}^n$ bilden, ist so nach (5.1.4) der Beweis bereits abgeschlossen.

(5.2.14) **Folgerung**

(i) Sind die Nullstellen des Polynoms P sämtlich einfach, d.h. $m = n$, $q_\mu = 1$ für $\mu = 1, \ldots, n$, so wird mit $\beta_\mu \in \mathbb{C}$

$$\eta_k = \sum_{\mu=1}^{n} \beta_\mu \lambda_\mu^k \qquad (k \in \mathbb{N}) .$$

(ii) Ist $a_0 \neq 0$, also $P(0) \neq 0$, so hat jede Lösung $(\eta_k)_0^\infty$ von (5.2.6) mit (von k unabhängigen) $\gamma_{\mu,\kappa} \in \mathbb{C}$ die Form

$$\eta_k = \sum_{\mu=1}^{m} \left(\sum_{\kappa=0}^{q_\mu - 1} \gamma_{\mu,\kappa} k^\kappa \lambda_\mu^k \right) \qquad (k \in \mathbb{N}) .$$

Im Fall $a_0 = 0$ gilt diese Darstellung wenigstens für alle $k \geq n$.

Abschließend behandeln wir in diesem Zusammenhang das *Verfahren von Bernoulli*. Hierzu gehen wir aus von einem Polynom

$$P(\lambda) = \lambda^n + a_{n-1} \lambda^{n-1} + \ldots + a_0 , \quad a_0 \neq 0$$

und fassen dieses als charakteristisches Polynom einer entsprechenden Differenzengleichung (5.2.6) auf. Über die Matrix A aus (5.2.7) ist dieser eine Rekursion des Typs (5.1.5) zugeordnet; im Sinne dieser Zuordnung formulieren wir unter Berücksichtigung des Hilfssatzes (5.2.10) den

(5.2.15) **Satz.** *$\lambda_1, \lambda_2, \ldots, \lambda_m$ seien die (verschiedenen) Nullstellen des Polynoms* P, *dabei sei*

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_m| .$$

Wir setzen $\rho = \left|\frac{\lambda_2}{\lambda_1}\right|$.

Mit $h_1 \in E_1, \neq 0$, $h_\mu \in H_\mu$ $(\mu = 2, \ldots, m)$ *gelte für den Startvektor* $y_0 = (\eta_k)_0^{n-1}$ *der Lösungsfolge* $(\eta_k)_0^\infty$ *von* (5.2.6) *die Zerlegung*

(5.2.16) $\quad y_0 = h_1 + h_2 + \ldots + h_m$.

Hierin seien die h_μ *Hauptvektoren der Stufe* p_μ *und damit*

$$p = \max\{p_\mu : 1 \leqq \mu \leqq m\} .$$

Unter diesen Voraussetzungen hat man

$$\lim_{k \to \infty} \frac{\eta_{k+1}}{\eta_k} = \lambda_1$$

bzw. genauer

$$\frac{\eta_{k+1}}{\eta_k} = \lambda_1 + O(\rho^k k^{p-1}) .$$

Zum *Beweis* stützen wir uns auf den Satz (5.1.10), (ii): nach Hilfssatz (5.2.10) ist $\dim E_1 = 1$ und daher mit einem $\alpha \in \mathbb{C}, \neq 0$ $h_1 = \alpha(\lambda_1^k)_0^{n-1}$, mithin insbesondere die erste Komponente des Vektors h_1 von Null verschieden. Zu beachten bleibt, daß η_k und η_{k+1} gemäß (5.2.2) die ersten Komponenten von y_k und y_{k+1} sind.

(5.2.17) **Zusatz.** Zur geeigneten Festlegung von $y_0 = (\eta_k)_0^{n-1}$ diskutieren wir die folgenden Möglichkeiten:

(i) Wählt man $y_0 \neq 0$ beliebig, so entsteht – wie in Abschnitt 5.1 eingehend begründet – durch Rundungsfehler nach einigen Iterationen eine von Null verschiedene Komponente in H_1. Ist λ_1 einfache Nullstelle von P, so liegt diese Komponente natürlich in E_1.

(ii) Geht man von

$$\eta_0 = \eta_1 = \ldots = \eta_{n-2} = 0, \quad \eta_{n-1} = 1$$

aus, so ist in der Zerlegung (5.2.16) $h_1 \neq 0$ gewährleistet.

(iii) Setzt man

$$\begin{cases} \eta_0 = -a_{n-1} , \\ \eta_\nu = -[(\nu+1)a_{n-\nu-1} + a_{n-\nu}\eta_0 + \ldots + a_{n-1}\eta_{\nu-1}] \qquad (\nu = 1, \ldots, n-1) , \end{cases}$$

so folgt

$$y_0 = \sum_{\mu=1}^{m} q_\mu \lambda_\mu (\lambda_\mu^k)_{k=0}^{n-1} ,$$

wonach die h_μ in (5.2.16) ausnahmslos Eigenvektoren sind.

Die Beweise zu (5.2.17), (ii) und (iii) werden als Übungsaufgaben 5.2 und 5.3 empfohlen. In der Aufgabe 5.3 ist eine Erweiterung des Bernoulli-Verfahrens beschrieben, mit der zugleich das reduzierte Polynom $(\lambda - \lambda_1)^{-1} P(\lambda)$ näherungsweise bestimmt wird. Hierauf beruht ein schneller konvergentes Verfahren zur Nullstellenberechnung von Polynomen, das von Jenkins und Traub [32], [33], [55] angegeben wurde. Das Bernoulli-Verfahren selbst hat wegen seiner i.a. langsamen Konvergenz kaum noch praktische Bedeutung.

5.3. Transformation auf obere Hessenbergform

Es sei $B = (b_{i,j})_{(n,n)} \in M(n \times n, \mathbb{C})$. Wir nennen B eine *Matrix in oberer Hessenbergform* oder auch kurz eine obere *Hessenbergmatrix*, wenn für alle $i > j + 1$ $b_{i,j} = 0$, d.h.

$$B = \begin{pmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,n} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,n} \\ 0 & b_{3,2} & \cdots & b_{3,n} \\ \vdots & \ddots & & \vdots \\ 0 & \cdots & 0\ b_{n,n-1} & b_{n,n} \end{pmatrix}$$

gilt.

Beabsichtigt ist, in diesem Abschnitt Konstruktionsverfahren anzugeben, die Ähnlichkeitstransformationen beliebiger (n,n)-Matrizen auf obere Hessenbergform bedeuten. Als erste Möglichkeit diskutieren wir eine derartige Transformation, die mit der in Abschnitt 2.2 dargestellten Gauß-Zerlegung eng verwandt ist und dementsprechend als Produkt elementarer unterer Dreiecksmatrizen und Permutationsmatrizen konstruiert wird.

Unter Verwendung der Bezeichnungen aus (2.2.2) beweisen wir zunächst in Ergänzung zu Hilfssatz (2.2.3) den

(5.3.1) **Hilfssatz.** *Ist* $C(m,n)$*-Matrix mit den Spalten* c_μ $(\mu = 1, \ldots, n)$, *so gilt für die Spalten* $\tilde{c}_\mu$ *von* $\tilde{C} = CL_\nu(d)$:

$$\tilde{c}_\mu = \begin{cases} c_\mu & \text{für } \mu \neq \nu, \\ c_\nu - \sum_{i=\nu+1}^{n} \delta_i c_i & \text{für } \mu = \nu. \end{cases}$$

Beweis. Man hat

$$L_\nu(d)\, e_\mu = e_\mu - d e_\nu^t e_\mu = \begin{cases} e_\mu & \text{für } \mu \neq \nu, \\ e_\nu - d = e_\nu - \sum_{i=\nu+1}^{n} \delta_i e_i & \text{für } \mu = \nu, \end{cases}$$

woraus wegen $\tilde{c}_\mu = C L_\nu(d)\, e_\mu$ unmittelbar die Behauptung folgt.

Wir notieren nun den

(5.3.2) **Satz** *(Transformation auf obere Hessenbergform). Es sei* A *eine reelle oder komplexe* (n,n)*-Matrix. Dann existiert eine invertierbare reelle bzw. komplexe* (n,n)*-Matrix* T, *so daß* $B = T^{-1}AT$ *obere Hessenbergform besitzt.*

Den *Beweis* verbinden wir mit einem *konstruktiven Verfahren* zur Berechnung von B und T. Da jede (2,2)-Matrix obere Hessenbergmatrix ist, nehmen wir o.E. $n \geqq 3$ an. Wir definieren

$$A^{(1)} := A = (a^{(1)}_{i,j})_{(n,n)} .$$

Beim 1. Konstruktionsschritt betrachten wir die folgenden beiden Fälle:

(I) Für alle $i = 2, \ldots, n$ gilt $a^{(1)}_{i,1} = 0$. Wir setzen $A^{(2)} = A^{(1)}$ und erhalten natürlich unmittelbar

$$(5.3.3) \qquad A^{(2)} = \begin{pmatrix} a^{(2)}_{1,1} & a^{(2)}_{1,2} & \cdots & a^{(2)}_{1,n} \\ a^{(2)}_{2,1} & a^{(2)}_{2,2} & \cdots & a^{(2)}_{2,n} \\ 0 & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & a^{(2)}_{n,2} & \cdots & a^{(2)}_{n,n} \end{pmatrix} .$$

(II) Es existiert ein $i \in \{2, \ldots, n\}$ mit $a^{(1)}_{i,1} \neq 0$. Wir wählen einen Index $i_2 \in \{2, \ldots, n\}$, so daß

$$|a^{(1)}_{i_2,1}| = \max_{i=2}^{n} |a^{(1)}_{i,1}| \qquad (> 0) .$$

Mit der elementaren Permutationsmatrix

$$P_{2,i_2} = I - (e_2 - e_{i_2})(e_2 - e_{i_2})^t$$

bilden wir zunächst

$$\tilde{A}^{(1)} := P_{2,i_2} A^{(1)} P_{2,i_2} = (\tilde{a}^{(1)}_{i,j})_{(n,n)} ,$$

worin $\tilde{a}^{(1)}_{2,1} = a^{(1)}_{i_2,1} \neq 0$ gilt, und hiermit weiter der Reihe nach

$$d_2 := \frac{1}{\tilde{a}^{(1)}_{2,1}} \sum_{\nu=3}^{n} \tilde{a}^{(1)}_{\nu,1} e_\nu$$

und

$$A^{(2)} := L_2(d_2)\tilde{A}^{(1)} L_2(-d_2) .$$

Wir beachten, daß nach Hilfssatz (2.2.3), (v) beim Übergang

$$(5.3.4) \qquad \tilde{A}^{(1)} \to L_2(d_2)\tilde{A}^{(1)}$$

die beiden ersten Zeilen von $\tilde{A}^{(1)}$ unverändert bleiben, während von den übrigen Zeilen mit den Indizes $i = 3, \ldots, n$ das $\frac{\tilde{a}^{(1)}_{i,1}}{\tilde{a}^{(1)}_{2,1}}$-fache der 2. Zeile abgezogen wird, wodurch in der 1. Spalte Nullen ab der 3. Komponente entstehen.

Weiter stellen wir fest, daß nach dem vorangehenden Hilfssatz (5.3.1) bei der Transformation

$$(5.3.5) \qquad L_2(d_2)\tilde{A}^{(1)} \to L_2(d_2)\tilde{A}^{(1)}L_2(-d_2) = A^{(2)}$$

nur die 2. Spalte verändert wird. Zu ihr wird nämlich die Summe des $\frac{\tilde{a}^{(1)}_{i,1}}{\tilde{a}^{(1)}_{2,1}}$-fachen der $i = 3, \ldots,$ n-ten Spalte addiert. Insgesamt ergibt sich daher, daß $A^{(2)}$ die Form (5.3.3) besitzt.

Ein 2. Konstruktionsschritt ist nur für $n \geq 4$ durchzuführen. Gilt $a^{(2)}_{i,2} = 0$ für $i = 3, \ldots, n$, so setzen wir $A^{(3)} = A^{(2)}$; ansonsten wählen wir ein $i_3 \in \{3, \ldots, n\}$ mit der Eigenschaft

$$|a^{(2)}_{i_3,2}| = \max_{i=3}^{n} |a^{(2)}_{i,2}| .$$

Hiervon ausgehend, bilden wir die Matrix

$$\tilde{A}^{(2)} := P_{3,i_3} A^{(1)} P_{3,i_3} = (\tilde{a}^{(2)}_{i,j})_{(n,n)} .$$

die ebenfalls die Gestalt (5.3.3) besitzt, und anschließend mit

$$d_3 := \frac{1}{\tilde{a}^{(2)}_{3,2}} \sum_{\nu=4}^{n} \tilde{a}^{(2)}_{\nu,2} e_\nu$$

weiter

$$A^{(3)} := L_3(d_3)\tilde{A}^{(2)}L_3(-d_3) .$$

Wie man sich analog zu (5.3.4), (5.3.5) überlegt, hat $A^{(3)}$ die Form

$$A^{(3)} = \begin{pmatrix} a^{(3)}_{1,1} & a^{(3)}_{1,2} & \cdots\cdots & a^{(3)}_{1,n} \\ a^{(3)}_{2,1} & a^{(3)}_{2,2} & \cdots\cdots & a^{(3)}_{2,n} \\ 0 & a^{(3)}_{3,2} & a^{(3)}_{3,3} & \vdots \\ \vdots & 0 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & a^{(3)}_{n,3} \cdots & a^{(3)}_{n,n} \end{pmatrix} .$$

Fährt man in dieser Weise fort, so erhält man nach $(n-2)$ Schritten mit

$$(5.3.6) \qquad T := P_{2,i_2} L_2(-d_2) \dots P_{n-1,i_{n-1}} L_{n-1}(-d_{n-1}) ,$$

worin gegebenenfalls für gewisse ν zusätzlich die P_{ν,i_ν} und $L_\nu(-d_\nu)$ als (n,n)-Einheitsmatrizen zu definieren sind, unter Berücksichtigung der Beziehungen (2.2.3), (iv) und (2.2.5), (vii), wie behauptet, $B = T^{-1}AT$ mit B in oberer Hessenbergform.

Wir ermitteln nun den *Rechenaufwand* dieses Verfahrens: die Transformation (5.3.4) benötigt $(n-2)$ Divisionen sowie $(n-1)(n-2)$ Multiplikationen; der Übergang (5.3.5) erfordert $n(n-2)$ Multiplikationen. Allgemein braucht man dementsprechend beim ν-ten Schritt

$(n-\nu+1)(n-\nu-1)$ Operationen gemäß (5.3.4),
$n(n-\nu-1)$ Operationen gemäß (5.3.5).

Summation über ν liefert

$$\sum_{\nu=1}^{n-2} (n+2+(n-\nu-1))(n-\nu-1) = (n+2)\sum_{j=1}^{n-2} j + \sum_{j=1}^{n-2} j^2 ,$$

folglich als Ergebnis die

(5.3.7) **Bemerkung.** Die Transformation einer (n,n)-Matrix auf obere Hessenbergform mittels elementarer Dreiecksmatrizen erfordert in etwa

$$\frac{5}{6} n^3 \text{ Multiplikationen bzw. Divisionen.}$$

Zur Rücktransformation von Eigenvektoren benötigt man nach den einführenden Überlegungen zu diesem Kapitel eine explizite Darstellung der Matrix T. Zu ihrer Herleitung beachtet man die Beziehung

$$(5.3.8) \qquad L_\nu(d_\nu) P_{i,j} = P_{i,j} L_\nu(P_{i,j} d_\nu) \qquad (i,j > \nu) ,$$

die wegen (2.2.5), (vii) unmittelbar aus (2.2.11) folgt. Nach (5.3.6) ergibt sich somit

$$(5.3.9) \qquad T = (P_{2,i} \dots P_{n-1,i_{n-1}})(L_2(-\hat{d}_2) \dots L_{n-1}(-\hat{d}_{n-1})) =: P^{-1}L ,$$

wobei die $\hat{d}_\nu$ und L ähnlich wie in (2.2.13) definiert sind.

Beim Programmieren beschreibt man die Zeilen- und Spaltenvertauschungen durch Permutationen $\pi_\nu \in S_n$; ferner speichert man die Koordinaten der d_ν an die Stellen der in den $A^{(\nu)}$ entstehenden Nullen. Auf diese Weise erhält man einen Algorithmus des Typs (2.2.16), dessen genaue Formulierung wir dem Leser als Übungsaufgabe 5.4 empfehlen. Jedenfalls sind zur Bestimmung von T keine zusätzlichen Rechenoperationen nötig.

Durch Rundungsfehler bedingt, gilt für die numerisch berechneten Matrizen $\tilde{T}$ und $\tilde{B}$ mit einer Störmatrix ΔA, die man mit den Methoden des Abschnitts 3.6 abschätzen kann, die Gleichung

$$\tilde{B} = \tilde{T}^{-1}(A + \Delta A)\tilde{T}. \tag{5.3.10}$$

Über die numerische Stabilität des vorstehend beschriebenen Verfahrens gelten infolgedessen ähnliche Aussagen wie für die Gauß-Elimination bei halbmaximaler Pivotwahl; Einzelheiten hierzu findet man bei Wilkinson [61], S. 363 ff. Der Einfluß der Störmatrix ΔA auf die Eigenwerte und Eigenvektoren der Matrix A bzw. B wird im Abschnitt 5.7 diskutiert.

Als Variante zu (5.3.2) notieren wir den

(5.3.11) **Satz** *(Transformation auf obere Hessenbergform nach Householder). Es sei* A *eine relle oder komplexe* (n,n)*-Matrix. Dann existiert eine unitäre reelle bzw. komplexe* (n,n)*-Matrix* Q, *so daß* $C = Q^*AQ$ *obere Hessenbergform besitzt.*

Beweis. Wir werden ein Konstruktionsverfahren angeben, das Q als Produkt von Householder-Matrizen liefert; im reellen Fall ist nach Givens – vgl. Aufgabe 2.11 – auch eine Darstellung mittels ebener Drehungen möglich.

Wie im Beweis zu (5.3.2) nehmen wir o.E. $n \geq 3$ an. Wir bezeichnen

$$A^{(1)} := A = (a_{i,j})_{(n,n)}$$

und unterteilen

$$A^{(1)} = \left(\begin{array}{c|c} c_{1,1} & B_1^* \\ \hline f_1 & E_1 \end{array}\right)$$

mit $c_{1,1} = a_{1,1}$, $B_1 = (\overline{a_{1,j}})_2^n$ und $f_1 = (a_{i,1})_2^n$. Ist für $i = 3, \ldots, n$ $a_{i,1} = 0$, so setzen wir $H_1 = I$, $A^{(2)} = A^{(1)}$; andernfalls wählen wir ähnlich wie im Beweis zu (2.6.8)

$$\left\{\begin{array}{l} \sigma_1 \in \mathbb{C} \text{ mit } |\sigma_1| = 1, \qquad \overline{\sigma_1} \cdot a_{2,1} = -|a_{2,1}|, \\ \mu_1 := |f_1| = \left(\sum_{i=2}^{n} |a_{i,1}|^2\right)^{1/2}, \\ \tilde{w}_1 := \dfrac{1}{|f_1 - \sigma_1 \mu_1 e_1|}(f_1 - \sigma_1 \mu_1 e_1), \end{array}\right. \tag{5.3.12}$$

wobei natürlich $e_1 \in \mathbb{C}^{n-1}$ sei. Mit $w_1 = \begin{pmatrix} 0 \\ \tilde{w}_1 \end{pmatrix} \in \mathbb{C}^n$ definieren wir

$$H_1 := H(w_1) = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & G_1 \end{array}\right) = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & I - 2\tilde{w}_1 \tilde{w}_1^* \end{array}\right).$$

Wir erhalten

$$H_1 A^{(1)} = \left(\begin{array}{c|c} c_{1,1} & B_1^* \\ \hline G_1 f_1 & G_1 E_1 \end{array}\right) = \left(\begin{array}{c|c} c_{1,1} & B_1^* \\ \hline -\sigma_1\mu_1 & \\ 0 & \\ \vdots & G_1 E_1 \\ 0 & \end{array}\right)$$

und weiter

$$A^{(2)} := H_1 A^{(1)} H_1 = \left(\begin{array}{c|c} c_{1,1} & B_1^* G_1 \\ \hline -\sigma_1\mu_1 & \\ 0 & \\ \vdots & G_1 E_1 G_1 \\ 0 & \end{array}\right).$$

Im Fall $n \geqq 4$ führen wir das Verfahren fort und unterteilen

$$A^{(2)} = \left(\begin{array}{cc|c} c_{1,1} & c_{1,2} & B_2^* \\ c_{2,1} & c_{2,2} & \\ \hline 0 & f_2 & E_2 \end{array}\right).$$

Verschwinden für $i = 4, \ldots, n$ sämtliche $a_{i,2}^{(2)}$, so setzen wir $H_2 = I$, $A^{(3)} = A^{(2)}$; ansonsten konstruieren wir zu f_2 gemäß der Vorschrift (5.3.12) ein $\tilde{w}_2 \in \mathbb{C}^{n-2}$. Zu

$$w_2 = \begin{pmatrix} 0 \\ 0 \\ \tilde{w}_2 \end{pmatrix} \in \mathbb{C}^n$$

betrachten wir

$$H_2 := H(w_2) = \left(\begin{array}{c|c} I_2 & 0 \\ \hline 0 & G_2 \end{array}\right)$$

und berechnen hiermit ähnlich wie beim 1. Schritt

$$A^{(3)} := H_2 A^{(2)} H_2 = \left(\begin{array}{cc|c} c_{1,1} & c_{1,2} & B_2^* G_2 \\ c_{2,1} & c_{2,2} & \\ \hline 0 & c_{3,2} & \\ 0 & 0 & G_2 E_2 G_2 \end{array}\right).$$

Nach $(n-2)$ Schritten dieser Art wird

$$C := A^{(n-1)} = (H_{n-2} H_{n-3} \dots H_1)\, A (H_1 \dots H_{n-3} H_{n-2})$$

obere Hessenbergmatrix; damit ist auf Grund der Bemerkung (2.6.5), (iii) die gewünschte Transformation ermittelt.

Daß bei dieser Konstruktion der Algorithmus (2.6.15) verwendet werden kann, zeigt die

(5.3.13) **Bemerkung**

(i) Die Berechnung von $H_1 A^{(1)}$ entspricht dem Übergang

$$(f_1, E_1) \to G_1 (f_1, E_1)$$

mit der durch f_1 bestimmten Householdermatrix G_1, bedeutet also die entsprechende Anwendung des Rechenschritts (2.6.15), (v).

(ii) Die sich anschließende Multiplikation $(H_1 A^{(1)}) H_1$ besteht in der Transformation

$$(B_1, (G_1 E_1)^*) \to G_1 (B_1, (G_1 E_1)^*),$$

mithin in der Ausführung eines Rechenschritts der Art (2.6.15), (vi).

Zur Ermittlung des *Rechenaufwandes* (im reellen Fall) stellen wir analoge Überlegungen wie zu (2.6.16) an. Die Ausführung von (5.3.13), (i) erfordert $2(n-1)n+1$, beim ν-ten Schritt entsprechend $2(n-\nu)(n-\nu+1)+1$ Multiplikationen bzw. Divisionen und je eine Wurzelberechnung. Zu (5.3.13), (ii) sind weitere $2n(n-1)$ bzw. beim ν-ten Schritt $2n(n-\nu)$ Multiplikationen notwendig. Aufsummieren ergibt

$$\sum_{\nu=1}^{n-2} [2(n-\nu)(n-\nu+1) + 2n(n-\nu)] = \sum_{j=2}^{n-1} 2j(j+1) + 2n \sum_{j=2}^{n-1} j$$

$$= \tfrac{1}{3}(n-1)\, n(5n+2) - 2n - 4$$

folglich zusammenfassend die

(5.3.14) **Bemerkung.** Die Transformation einer (n,n)-Matrix auf obere Hessenbergform mittels Householder-Matrizen erfordert $(n-2)$ Wurzelberechnungen sowie

$$\approx \frac{5}{3} n^3 \text{ Multiplikationen bzw. Divisionen.}$$

Eine Rundungsfehleranalyse liefert ein zu (5.3.10) analoges Ergebnis; sie ist mit den Abschätzungstechniken des Abschnitts 3.7 durchzuführen.

Vergleicht man die beiden in den Sätzen (5.3.2) und (5.3.11) angegebenen Transformationsmethoden, so kommt man zu einem ähnlichen Ergebnis wie bei der Gegenüberstellung der Gauß- und der Householder-Elimination: Wegen ihres geringeren Rechenaufwandes ist im allgemeinen die Transformation nach Gauß derjenigen nach Householder vorzuziehen.

Eine Ausnahme bilden jedoch die hermiteschen Matrizen; diese wird man in der Regel nach Householder, also mit unitären Matrizen transformieren. Zur Begründung notieren wir die (elementar beweisbare)

(5.3.15) **Bemerkung.** Ist A hermitesch, Q unitär und $C = Q^*AQ$, so ist auch C hermitesch. Eine hermitesche obere Hessenbergmatrix $C = (c_{i,j})_{(n,n)}$ hat Tridiagonalgestalt, nämlich die Eigenschaft, daß die $c_{i,j}$ für Indizes mit $|i-j| \geq 2$ verschwinden.

Im Fall einer hermiteschen Matrix A sind übrigens sämtliche bei der Konstruktion gemäß Satz (5.3.11) auftretenden $A^{(\nu)} = (a_{i,j}^{(\nu)})_{(n,n)}$ hermitesch. Aus diesem Grunde hat man nur die Koeffizienten $a_{i,j}^{(\nu)}$ für $i \geq j$ zu berechnen; ein dazu geeigneter Algorithmus ist in der Übungsaufgabe 5.5 vorgeschlagen. Dementsprechend kann man nachweisen, daß die Householder-Reduktion einer hermiteschen Matrix auf Tridiagonalgestalt im wesentlichen $\frac{2}{3}n^3$ Rechenoperationen benötigt.

5.4. Das charakteristische Polynom einer oberen Hessenbergmatrix

Es sei $B = (b_{i,j})_{(n,n)}$ eine obere Hessenbergmatrix. Zunächst sind wir darauf aus, ein Verfahren anzugeben, das bei beliebig vorgegebenem $\lambda \in \mathbb{C}$ die Berechnung von $\varphi(\lambda) = \det(\lambda I - B)$ sowie $\varphi'(\lambda)$ gestattet. Dazu setzen wir o.E.

(5.4.1) $\quad b_{i+1,i} \neq 0 \qquad (i = 1, \ldots, n-1)$

voraus. Ist nämlich für ein festes j $b_{j+1,j} = 0$, so gilt

$$\det(\lambda I - B) = \det \begin{pmatrix} \lambda I_j - A_j & \\ 0 & \lambda I_{n-j} - C_{n-j} \end{pmatrix} = \det(\lambda I_j - A_j) \cdot \det(\lambda I_{n-j} - C_{n-j})$$

mit (j,j)- bzw. (n−j, n−j)-Matrizen A_j und C_{n-j} in oberer Hessenbergform; durch Übergang auf Teilmatrizen ist also eine Reduktion auf die Voraussetzung (5.4.1) möglich.

Definiert man gemäß Hyman [30] rekursiv

$$(5.4.2) \quad \begin{cases} x_n(\lambda) = 1\,, \\ x_{n-1}(\lambda) = \dfrac{1}{b_{n,n-1}} (\lambda - b_{n,n}) x_n(\lambda)\,, \\ x_{j-1}(\lambda) = \dfrac{1}{b_{j,j-1}} \left[(\lambda - b_{j,j}) x_j(\lambda) - \sum_{i=j+1}^{n} b_{j,i} x_i(\lambda) \right] \quad (j = n-1, \ldots, 2)\,, \end{cases}$$

so ist klar, daß die $x_i(\lambda)$ sämtlich Polynome in λ sind. Ferner rechnet man leicht nach, daß der Vektor $x(\lambda) = (x_i(\lambda))_1^n$ der Gleichung

(5.4.3) $\quad (\lambda I - B)\, x(\lambda) = c(\lambda) e_1$

genügt, worin

$$(5.4.4) \qquad c(\lambda) = (\lambda - b_{1,1})x_1(\lambda) - \sum_{i=2}^{n} b_{1,i}x_i(\lambda)$$

zu setzen ist. Weiter zeigen wir

(5.4.5) **Satz.** *Mit*

$$\frac{1}{\alpha} := \prod_{i=1}^{n-1} b_{i+1,i}$$

gilt für alle $\lambda \in \mathbb{C}$

$$c(\lambda) = \alpha \cdot \varphi(\lambda) = \alpha \cdot \det(\lambda I - B) .$$

Beweis. Wir definieren zu $x(\lambda)$ die (n,n)-Matrix

$$R(\lambda) := \begin{pmatrix} 1 & 0 & \dots & 0 & x_1(\lambda) \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ \vdots & & \ddots & 1 & x_{n-1}(\lambda) \\ 0 & & & 0 & x_n(\lambda) \end{pmatrix}$$

und erhalten wegen $\det R(\lambda) = x_n(\lambda) = 1$ unmittelbar die Gleichung

$$(5.4.7) \qquad \det(\lambda I - B) = \det[(\lambda I - B)\, R(\lambda)] .$$

Wegen

$$R(\lambda)\, e_j = \begin{cases} e_j & (j = 1, \dots, n-1) , \\ x(\lambda) & (j = n) \end{cases}$$

ergibt sich unter Berücksichtigung von (5.4.3)

$$(\lambda I - B)\, R(\lambda) = \begin{pmatrix} \lambda - b_{1,1} & \dots & \dots & -b_{1,n-1} & c(\lambda) \\ -b_{2,1} & \ddots & & \vdots & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \lambda - b_{n-1,n-1} & \vdots \\ 0 & \dots & 0 & -b_{n,n-1} & 0 \end{pmatrix} .$$

Die Determinante dieser Matrix berechnen wir durch Entwickeln nach der letzten Spalte zu

$$\det[(\lambda I - B)\, R(\lambda)] = \left(\prod_{i=1}^{n-1} b_{i+1,i} \right) \cdot c(\lambda) ,$$

und ein Vergleich mit (5.4.7) liefert die Behauptung des Satzes.

Zur Bestimmung einer Nullstelle von $\varphi(\lambda)$ bzw. $c(\lambda)$ kann das im 6. Kapitel beschriebene Newton-Verfahren

$$\lambda^{(k+1)} = \lambda^{(k)} - \frac{\varphi(\lambda^{(k)})}{\varphi'(\lambda^{(k)})} \qquad (k = 0, 1, 2, \ldots),$$

wobei $\lambda^{(0)}$ eine geeigente Ausgangsnäherung ist, angewendet werden. Hierzu muß neben $c(\lambda) = \alpha \cdot \varphi(\lambda)$ die Ableitung $c'(\lambda) = \alpha \cdot \varphi'(\lambda)$ für i.a. mehrere $\lambda \in \mathbb{C}$ ermittelt werden. Ein rekursives Verfahren zu ihrer Berechnung ergibt sich aus (5.4.2), (5.4.4) durch Differentiation; auf diese Weise erhält man nämlich neben $x_n'(\lambda) = 0$ die Rekursionen

$$(5.4.8)\quad \begin{cases} x_{n-1}'(\lambda) = \dfrac{1}{b_{n,n-1}} x_n(\lambda), \\ x_{j-1}'(\lambda) = \dfrac{1}{b_{j,j-1}} \left[x_j(\lambda) + (\lambda - b_{j,j}) x_j'(\lambda) - \displaystyle\sum_{i=j+1}^{n-1} b_{j,i} x_i'(\lambda) \right] \\ \qquad\qquad (j = n-1, \ldots, 2), \\ c'(\lambda) = x_1(\lambda) + (\lambda - b_{1,1}) x_1'(\lambda) - \displaystyle\sum_{i=2}^{n-1} b_{1,i} x_i'(\lambda), \end{cases}$$

worin die auftretenden $x_j(\lambda)$ bereits im Zusammenhang mit der Berechnung von $c(\lambda)$ bestimmt sind.

Voraussetzungen für die Konvergenz des Newton-Verfahrens werden im 6. Kapitel angegeben; u.a. wird gefordert, daß $\lambda^{(0)}$ der gesuchten Nullstelle genügend nahe liegt. Ist das Newton-Verfahren nicht anwendbar, so sind die Methoden, die gezielt zur Berechnung der Nullstellen von Polynomen entwickelt worden sind, wie z.B. die Verfahren von Nickel [42], Jenkins und Traub [32], [33] heranzuziehen.

In dem wichtigen Spezialfall einer hermiteschen oberen Hessenbergmatrix, d.h. im Fall einer *hermiteschen Tridiagonalmatrix* sind wir in der Lage, die Nullstellenberechnung von $\varphi(\lambda)$ hier vollständig zu behandeln. Hierzu beweisen wir zunächst den

(5.4.9) **Hilfssatz.** *Es sei* C *eine hermitesche Tridiagonalmatrix, also mit* $\alpha_i \in \mathbb{C}$ *und* $\beta_i \in \mathbb{R}$

$$C = \begin{pmatrix} \beta_1 & \bar{\alpha}_1 & & & 0 \\ \alpha_1 & \beta_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \bar{\alpha}_{n-1} \\ 0 & & & \alpha_{n-1} & \beta_n \end{pmatrix}.$$

Dann genügt die Folge der k*-zeiligen Hauptunterdeterminanten* $q_k(\lambda)$ *von* $q_n(\lambda) = \det(\lambda I - C)$ *(mit der üblichen zusätzlichen Festlegung* $q_0(\lambda) := 1$*) der linearen Rekursion*

$$(5.4.10)\quad \begin{cases} q_0(\lambda) = 1,\ q_1(\lambda) = \lambda - \beta_1\ , \\ q_{k+1}(\lambda) = (\lambda - \beta_{k+1})\, q_k(\lambda) - |\alpha_k|^2 q_{k-1}(\lambda) \qquad (k = 1, \ldots, n-1) \end{cases}$$

und ist durch diese bestimmt.

Beweis. Die ersten beiden Gleichungen in (5.4.10) gelten trivialerweise. Die Rekursion beweist man, indem man die Determinante $q_{k+1}(\lambda)$ nach der letzten Zeile entwickelt. Es ergibt sich

$$q_{k+1}(\lambda) = (\lambda - \beta_{k+1})\, q_k(\lambda) - \alpha_k \cdot \det \begin{pmatrix} \lambda - \beta_1 & \bar{\alpha}_1 & & 0 & 0 \\ \alpha_1 & & \ddots & & \vdots \\ 0 & \ddots & & \bar{\alpha}_{k-2} & \vdots \\ \vdots & & \lambda - \beta_{k-1} & & 0 \\ 0 & \cdots\cdots & 0 & \alpha_{k-1} & \bar{\alpha}_k \end{pmatrix},$$

folglich die Behauptung, da die letzte Determinante den Wert $\bar{\alpha}_k\, q_{k-1}(\lambda)$ hat.

Entsprechend (5.4.1) setzen wir für das folgende

$$(5.4.11)\quad \alpha_i \neq 0 \qquad (i = 1, \ldots, n-1)$$

voraus. Dann bilden die $q_k(\lambda)$ eine sogenannte *Sturmsche Kette*; über ihre Nullstellenverteilung berichtet der

(5.4.12) **Satz**

(i) *Jedes* q_k *besitzt genau* k *verschiedene reelle Nullstellen*

$$\lambda_k^{(k)} < \lambda_{k-1}^{(k)} < \ldots < \lambda_1^{(k)}\ .$$

(ii) *Setzt man für* $k = 0, \ldots, n$ *zusätzlich*

$$\lambda_{k+1}^{(k)} := -\infty\ ,\quad \lambda_0^{(k)} := +\infty\ ,$$

so ergibt sich für $j = 1, \ldots, k$

$$\lambda_j^{(k-1)} < \lambda_j^{(k)} < \lambda_{j-1}^{(k-1)}\ .$$

Wir führen den *Beweis* durch Induktion über k. $q_1(\lambda) = \lambda - \beta_1$ hat genau eine reelle Nullstelle $\lambda_1^{(1)} = \beta_1$, und es gilt trivialerweise

$$-\infty = \lambda_1^{(0)} < \lambda_1^{(1)} < \lambda_0^{(0)} = +\infty\ .$$

Weiter nehmen wir an, die Behauptung des Satzes gelte bis zu einem k, wobei $1 \leqq k \leqq n-1$ sei. Mit (5.4.10), (5.4.11) folgt dann für $j = 1, \ldots, k$

$$q_{k+1}(\lambda_j^{(k)}) = -|\alpha_k|^2 q_{k-1}(\lambda_j^{(k)}) \neq 0$$

und daher

$$(5.4.13) \qquad q_{k+1}(\lambda_j^{(k)})\, q_{k-1}(\lambda_j^{(k)}) = -|\alpha_k q_{k-1}(\lambda_j^{(k)})|^2 < 0 .$$

Als nächstes stellen wir fest, daß für $j = 1, \ldots, k$ die Implikation

$$(5.4.14) \qquad \lambda \in]\lambda_j^{(k-1)}, \lambda_{j-1}^{(k-1)}[\Rightarrow (-1)^{j-1} q_{k-1}(\lambda) > 0$$

zutrifft. Das Polynom q_{k-1} wechselt nämlich innerhalb der angegebenen Intervalle das Vorzeichen nicht, macht dies jedoch in jeder der allesamt einfachen Nullstellen $\lambda_1^{(k-1)}, \lambda_2^{(k-1)}, \ldots, \lambda_{k-1}^{(k-1)}$. Es bleibt dann nur noch

$$q_{k-1}(\lambda) \to +\infty \qquad (\lambda \to +\infty)$$

zu beachten. Nach Induktionsannahme liegt $\lambda = \lambda_j^{(k)}$ in $]\lambda_j^{(k-1)}, \lambda_{j-1}^{(k-1)}[$; infolgedessen haben wir nach (5.4.14) für $j = 1, \ldots, k$

$$(-1)^{j-1} q_{k-1}(\lambda_j^{(k)}) > 0$$

und daher weiter wegen (5.4.13)

$$(-1)^j q_{k+1}(\lambda_j^{(k)}) > 0 .$$

Hiernach und wegen

$$q_{k+1}(\lambda) \to +\infty \ (\lambda \to +\infty), \ (-1)^{k+1} q_{k+1}(\lambda) \to +\infty \ (\lambda \to -\infty)$$

wechselt q_{k+1} beim Durchgang durch jedes der Intervalle $]\lambda_j^{(k)}, \lambda_{j-1}^{(k)}[$ $(j = 1, \ldots, k+1)$ das Vorzeichen; folglich besitzt q_{k+1} in jedem dieser $k+1$ Intervalle mindestens eine und mithin auch genau eine Nullstelle $\lambda_j^{(k+1)}$.

Für $1 \leqq k \leqq n$ und $\lambda \in \mathbb{R}$ bezeichnen wir mit $\nu_k(\lambda)$ die Anzahl der Vorzeichenwechsel in $(q_0(\lambda), q_1(\lambda), \ldots, q_k(\lambda))$. Dabei werden die reellen Zahlen $q_0(\lambda), \ldots, q_k(\lambda)$ in der angegebenen Reihenfolge durchlaufen und Nullen nicht berücksichtigt.

(5.4.15) **Satz.** *Für jedes* $\lambda \in \mathbb{R}$ *gibt* $\nu_n(\lambda)$ *die Anzahl der Nullstellen von* q_n, *die echt oberhalb* λ *liegen, an.*

Zum *Beweis* zeigen wir, daß für $0 \leqq j \leqq k$ und $1 \leqq k \leqq n$ die Aussage

$$(5.4.16) \qquad \nu_k(\lambda) = j \Rightarrow \lambda \in [\lambda_{j+1}^{(k)}, \lambda_j^{(k)}[$$

gilt.

Für $k = 1$ haben wir wegen $q_0(\lambda) = 1 \ (> 0)$

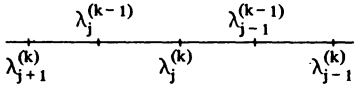
$$\begin{cases} \nu_1(\lambda) = 0 \Rightarrow q_1(\lambda) \geqq 0 \Rightarrow \lambda \geqq \lambda_1^{(1)}, \\ \nu_1(\lambda) = 1 \Rightarrow q_1(\lambda) < 0 \Rightarrow \lambda < \lambda_1^{(1)}. \end{cases}$$

Zum Schluß von $k-1$ auf k gehen wir von $\nu_k(\lambda) = j$ aus. Gemäß Definition tritt

entweder (I) $\nu_{k-1}(\lambda) = j - 1$ oder (II) $\nu_{k-1}(\lambda) = j$

ein. Wir diskutieren zunächst den Fall (I). Nach Induktionsvoraussetzung und (5.4.12), (ii) ergibt sich

$$\lambda \in [\lambda_j^{(k-1)}, \lambda_{j-1}^{(k-1)}[\subset]\lambda_{j+1}^{(k)}, \lambda_j^{(k)}[\cup [\lambda_j^{(k)}, \lambda_{j-1}^{(k)}[\,. \tag{5.4.17}$$



Wäre nun $\lambda = \lambda_j^{(k)}$, so hätte man $q_k(\lambda) = 0$ und damit den Widerspruch

$$\nu_k(\lambda) = \nu_{k-1}(\lambda) \,.$$

Ebenfalls zu diesem Widerspruch führt die Annahme $\lambda \in]\lambda_j^{(k)}, \lambda_{j-1}^{(k)}[$. Es wäre dann nämlich nach (5.4.14), angewendet bezüglich k,

$$(-1)^{j-1} q_k(\lambda) > 0 \,,$$

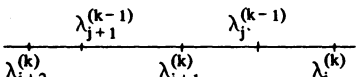
ferner $\lambda \neq \lambda_j^{(k-1)}$, mithin $\lambda \in]\lambda_j^{(k-1)}, \lambda_{j-1}^{(k-1)}[$ und daher abermals nach (5.4.14)

$$(-1)^{j-1} q_{k-1}(\lambda) > 0 \,.$$

Insgesamt gesehen, bleibt nach (5.4.17) nur die Möglichkeit $\lambda \in]\lambda_{j+1}^{(k)}, \lambda_j^{(k)}[$.

Im Fall (II) gilt nach Induktionsvoraussetzung und (5.4.12), (ii)

$$\lambda \in [\lambda_{j+1}^{(k-1)}, \lambda_j^{(k-1)}[\subset]\lambda_{j+2}^{(k)}, \lambda_{j+1}^{(k)}[\cup [\lambda_{j+1}^{(k)}, \lambda_j^{(k)}[\,. \tag{5.4.18}$$



Wäre nun $\lambda = \lambda_{j+1}^{(k-1)}$, so hätte man $q_{k-1}(\lambda) = 0$, ferner $j + 1 \leq k - 1$ sowie nach (5.4.13), angewendet bezüglich $k - 1$, $q_{k-2}(\lambda)\, q_k(\lambda) < 0$, mithin den Widerspruch

$$\nu_k(\lambda) = \nu_{k-1}(\lambda) + 1 \,.$$

Danach liegt λ also notwendigerweise in $]\lambda_{j+1}^{(k-1)}, \lambda_j^{(k-1)}[$. Dies führt – man beachte $j = \nu_{k-1}(\lambda) \leq k - 1$ – gemäß (5.4.14) zu

$$(-1)^j q_{k-1}(\lambda) > 0 \,.$$

Läge nun λ in $]\lambda_{j+2}^{(k)}, \lambda_{j+1}^{(k)}[$, so ergäbe sich abermals nach (5.4.14), diesmal bezüglich k benutzt,

$$(-1)^{j+1} q_k(\lambda) > 0$$

und damit der gleiche Widerspruch wie unter der Annahme $\lambda = \lambda_{j+1}^{(k-1)}$. Nach (5.4.18) ist damit auch hier die Behauptung bewiesen.

Zur Abkürzung setzen wir

$$\nu(\lambda) := \nu_n(\lambda), \qquad \lambda_i := \lambda_i^{(n)} \quad (i = 1, \ldots, n)$$

und notieren als

(5.4.19) **Folgerung.** Für $1 \leqq j \leqq n$ und $\lambda \in \mathbb{R}$ gilt

$$\nu(\lambda) \leqq j-1 \Rightarrow \lambda \geqq \lambda_j ,$$
$$\nu(\lambda) \geqq j \Rightarrow \lambda < \lambda_j .$$

Auf dieser Tatsache beruht das *Bisektionsverfahren;* es dient zur Berechnung der Nullstellen λ_j von q_n, d.h. der Eigenwerte von C:

Zunächst ermittelt man $a_0, b_0 \in \mathbb{R}$, so daß für $i = 1, \ldots, n$

$$a_0 < \lambda_i < b_0 .$$

Dazu stützt man sich z.B. auf den 1. Satz von Gerschgorin – siehe (5.7.5) – und wählt dementsprechend mit einem $\epsilon > 0$

$$a_0 := \min_{i=1}^{n} (\beta_i - |\alpha_i| - |\alpha_{i-1}|) - \epsilon ,$$
$$b_0 := \max_{i=1}^{n} (\beta_i + |\alpha_i| + |\alpha_{i-1}|) + \epsilon ,$$

worin $\alpha_0 = \alpha_n = 0$ zu setzen ist. Wegen $a_0 < \lambda_n$, $b_0 > \lambda_1$ gilt dann für jedes $j = 1, \ldots, n$

$$\nu(a_0) = n \geqq j , \quad \nu(b_0) = 0 \leqq j-1 .$$

Anschließend fixiert man ein derartiges j und ermittelt für $m = 0, 1, 2, \ldots$ rekursiv die Werte

$$(5.4.20) \quad \begin{cases} c_m := \frac{1}{2}(a_m + b_m) , \\ a_{m+1} := a_m , \quad b_{m+1} := c_m , \quad \text{falls } \nu(c_m) \leqq j-1 , \\ a_{m+1} := c_m , \quad b_{m+1} := b_m , \quad \text{falls } \nu(c_m) \geqq j . \end{cases}$$

Die Konvergenz dieses Verfahrens, d.h. die Konvergenz

$$a_m \nearrow \lambda_j , \; b_m \searrow \lambda_j \qquad (m \to \infty)$$

folgt unmittelbar daraus, daß gemäß Konstruktion für $m = 0, 1, 2, \ldots$

$$\nu(a_m) \geqq j , \; \nu(b_m) \leqq j-1 , \; b_{m+1} - a_{m+1} = \tfrac{1}{2}(b_m - a_m)$$

und daher nach (5.4.19)

$$\lambda_j \in [a_m, b_m] , \; b_m - a_m = \frac{1}{2^m}(b_0 - a_0)$$

gilt. Ferner erkennt man, daß bei jeder Iteration die Genauigkeit um eine Dualziffer verbessert wird. Ist λ_j von den benachbarten Nullstellen hinreichend getrennt, so wird für $m \geqq m_0$

$$\nu(a_m) = j , \; \nu(b_m) = j-1 ;$$

es liegt dann also in $[a_m, b_m]$ nur die Nullstelle λ_j. Ist dieser Fall eingetreten, so ist es ratsam, nach einigen weiteren Schritten auf das schneller konvergente Newton-Verfahren umzusteigen. Die hierzu benötigte Ableitung von q_n ermittelt man mittels einer Rekursion, die man durch Differentiation aus (5.4.10) herleitet.

(5.4.21) **Beispiel.** Wir betrachten die Matrix

$$A = \begin{pmatrix} 6 & 4 & 4 & 1 \\ 4 & 6 & 1 & 4 \\ 4 & 1 & 6 & 4 \\ 1 & 4 & 4 & 1 \end{pmatrix} ,$$

die, wie man leicht nachprüft, die Eigenwerte $\lambda_1 = 15$, $\lambda_2 = \lambda_3 = 5$, $\lambda_4 = -1$ besitzt. A wird gemäß Satz (5.3.11) in eine symmetrische Tridiagonalmatrix C transformiert. Bei 7-stelliger Gleitkommarechnung gewinnen wir als numerische Näherung

$$\tilde{C} = \begin{pmatrix} 6{,}000000 & -5{,}744563 & 0 & 0 \\ -5{,}744563 & 8{,}908946 & -5{,}142558 & 0 \\ 0 & -5{,}142558 & 4{,}090911 & 0{,}246 \cdot 10^{-4} \\ 0 & 0 & 0{,}246 \cdot 10^{-4} & 4{,}999979 \end{pmatrix} .$$

In der exakten Matrix C hat wegen $\lambda_2 = \lambda_3$ und Satz (5.4.12) einer der Koeffizienten α_i, und zwar offenbar α_3, den Wert Null. Zur Berechnung der Eigenwerte $\tilde{\lambda}_2$ und $\tilde{\lambda}_3$ von $\tilde{C}$ bestimmen wir mit dem Bisektionsverfahren Intervalle $[a_m, b_m]$ bzw. $[a'_m, b'_m]$ mit

$$\tilde{\lambda}_2 \in [a_m, b_m] , \ \tilde{\lambda}_3 \in [a'_m, b'_m] \qquad (m = 0, 1, 2, \ldots) .$$

Hierbei gewinnen wir, von

$$[a_0, b_0] = [a'_0, b'_0] := [-1{,}978186, 19{,}796068]$$

ausgehend, nach 18 Bisektionen

$$[a_{18}, b_{18}] = [a'_{18}, b'_{18}] = [4{,}999959, 5{,}000042], \ c_{18} = c'_{18} = 5{,}000001$$

mit

$$\nu(a_{18}) = 3, \ \nu(c_{18}) = 2, \ \nu(b_{18}) = 1$$

und weiter nach (5.4.20)

$$[a_{19}, b_{19}] = [c_{18}, b_{18}], \ [a'_{19}, b'_{19}] = [a_{18}, c_{18}] .$$

Da mithin $\tilde{\lambda}_2$ und $\tilde{\lambda}_3$ sehr dicht beieinander liegen, empfiehlt es sich nicht, das Newton-Verfahren zu verwenden; statt dessen erhalten wir nach je fünf weiteren Bisektionsschritten die Werte

$$\tilde{\lambda}_2 = 5{,}000004, \ \tilde{\lambda}_3 = 4{,}999967 .$$

5.5. QR- und LR-Verfahren

Das LR-Verfahren wurde 1958 von Rutishauser [47] angegeben; in Analogie dazu hat Francis [20] 1961 das QR-Verfahren entwickelt. Da das LR-Verfahren nicht immer anwendbar ist, wollen wir nur das QR-Verfahren ausführlicher darstellen.

Es sei $A \in M(n \times n, \mathbb{C})$; dann lautet der

(5.5.1) *Algorithmus des* QR-*Verfahrens:* Wir setzen

$$A_0 := A ;$$

für $k = 0, 1, 2, \ldots$ zerlegen wir

$$A_k = Q_k R_k ,$$

wobei die Q_k unitär, die R_k obere Dreiecksmatrizen sind, und definieren hiermit

$$A_{k+1} := R_k Q_k .$$

(5.5.2) **Bemerkungen.**

(i) $A_k = (Q_0 \ldots Q_{k-1})^* A (Q_0 \ldots Q_{k-1}) \qquad (k = 1, 2, 3, \ldots)$,

die A_k sind also sämtlich ähnlich zu A;

(ii) $A^k = (Q_0 \ldots Q_{k-1})(R_{k-1} \ldots R_0) \qquad (k = 1, 2, 3, \ldots)$.

Den *Beweis* von (i) führen wir durch Induktion. Zunächst haben wir

$$A_1 = R_0 Q_0 , \quad R_0 = Q_0^* A_0 ,$$

womit der Fall $k = 1$ erledigt ist. Für $k \geq 1$ gilt ebenso

$$A_{k+1} = R_k Q_k , \quad R_k = Q_k^* A_k ,$$

also $A_{k+1} = Q_k^* A_k Q_k$; unter der Annahme, daß (i) für k bereits gilt, folgern wir die Aussage für $k + 1$.

Die Gleichung (ii) für $k = 1$ ist klar; aus der Gültigkeit von (ii) für ein festes k folgern wir mit (i) und der Zerlegung von A_k

$$\begin{aligned} A^{k+1} &= A(Q_0 \ldots Q_{k-1})(R_{k-1} \ldots R_0) = (Q_0 \ldots Q_{k-1}) A_k (R_{k-1} \ldots R_0) \\ &= (Q_0 \ldots Q_{k-1}) Q_k R_k (R_{k-1} \ldots R_0) . \end{aligned}$$

Wir machen bezüglich A zunächst die

(5.5.3) **Voraussetzung.**

(i) Es existiere eine invertierbare (n,n)-Matrix T mit

$$T^{-1} A T = D := \operatorname{diag}(\lambda_1, \ldots, \lambda_n) ;$$

hierbei sei

$$|\lambda_1| \geqq |\lambda_2| \geqq \ldots \geqq |\lambda_n| > 0\,,\quad |\lambda_1| > |\lambda_n|\,;$$

(ii) T^{-1} besitze eine Zerlegung

$$T^{-1} = LU$$

in eine normierte untere Dreiecksmatrix L und eine obere Dreiecksmatrix U;

(iii) es gelte

$$|\lambda_i| = |\lambda_{i+1}| \Rightarrow \lambda_i = \lambda_{i+1} \qquad (i = 1, \ldots, n-1)\,.$$

Wir benutzen die

(5.5.4) **Bezeichnungen.** Es sei

$$\rho := \max\left\{\frac{|\lambda_i|}{|\lambda_j|}:\ i,j \in \{1,\ldots,n\},\quad |\lambda_i| < |\lambda_j|\right\};$$

$$\rho_n := \max\left\{\frac{|\lambda_n|}{|\lambda_j|}:\quad j \in \{1,\ldots,n-1\}, |\lambda_n| < |\lambda_j|\right\}.$$

Natürlich gilt $0 < \rho_n \leqq \rho < 1$. Als Normen im $\mathbb{C}^n$ (auch für (1,n)-Matrizen, also Zeilen) und $M(n \times n, \mathbb{C})$ wollen wir die euklidische bzw. die Frobeniusnorm – vgl. (3.3.8) – verwenden.

Mit diesen Voraussetzungen und Bezeichnungen beweisen wir den

(5.5.5) **Hilfssatz.** *Es existiert eine normierte untere Dreiecksmatrix* $\tilde{L} \in M(n \times n, \mathbb{C})$ *mit den Eigenschaften*

$$\begin{cases} \tilde{L}D = D\tilde{L}\,, \\ D^k L D^{-k} = \tilde{L}(I + F_k) \quad (k \in \mathbb{N}),\ \lim\limits_{k\to\infty} F_k = 0\,. \end{cases}$$

Hierbei besitzen die F_k *für* $k \to \infty$ *das Konvergenzverhalten*

$$|F_k| = O(\rho^k),\ |e_n^t F_k| = O(\rho_n^k)\,.$$

Beweis. Es sei $L =: (l_{i,j})_{(n,n)}$. Wir bezeichnen

$$M_k = (m_{i,j}^{(k)})_{(n,n)} := D^k L D^{-k} \quad (k \in \mathbb{N})\,;$$

die M_k sind untere Dreiecksmatrizen mit den Koeffizienten

$$m_{i,j}^{(k)} = l_{i,j}\left(\frac{\lambda_i}{\lambda_j}\right)^k.$$

Wir definieren $\tilde{L} = (\tilde{l}_{i,j})_{(n,n)}$ durch

$$(5.5.6) \qquad \tilde{l}_{i,j} = \begin{cases} l_{i,j}\,, & \text{falls } i \geqq j,\ \lambda_i = \lambda_j\,, \\ 0 & \text{sonst}\,. \end{cases}$$

Offenbar ist $\tilde{L}$ normierte untere Dreiecksmatrix mit $\tilde{L}D = D\tilde{L}$. Wenn wir

$$M_k =: \tilde{L} + \tilde{M}_k \quad (k \in \mathbb{N})$$

aufteilen, so erhalten wir auf Grund von (5.5.3), (iii)

$$\lim_{k \to \infty} \tilde{M}_k = 0\,, \quad \exists\, \tilde{\gamma} \geq 0 \quad |\tilde{M}_k| \leq \tilde{\gamma}\, \rho^k \quad (k \in \mathbb{N})\,.$$

Daher gelten mit

$$F_k := \tilde{L}^{-1} \tilde{M}_k \quad (k \in \mathbb{N})$$

natürlich die Darstellungen $M_k = \tilde{L}(I + F_k)$; außerdem folgt

$$|F_k| \leq \tilde{\gamma}\, |\tilde{L}^{-1}|\, \rho^k \quad (k \in \mathbb{N})$$

und damit $\lim_{k \to \infty} F_k = 0$, $|F_k| = O(\rho^k)$ $(k \to \infty)$ unmittelbar.

Nun zum Konvergenzverhalten der n-ten Zeilen der F_k. Mit der Bezeichnung

(5.5.7) $$s := \min \{i \in \{1, \ldots, n\} : |\lambda_i| = |\lambda_n|\} \quad (\geq 2)$$

wollen wir die Abschätzungen

(5.5.8) $$|e_i^t F_k| \leq \gamma_i \rho_n^k \qquad (i = s, \ldots, n;\ k \in \mathbb{N})$$

zeigen, wobei die $\gamma_i \geq 0$ geeignete Konstanten bedeuten.

Zunächst schließen wir aus

$$\tilde{L} F_k = \tilde{M}_k\,, \quad e_s^t \tilde{L} = e_s^t$$

die Beziehung

$$e_s^t F_k = e_s^t \tilde{L} F_k = e_s^t \tilde{M}_k \quad (k \in \mathbb{N})\,,$$

womit (5.5.8) für $i = s$ gezeigt ist. Im Fall $s < n$ nehmen wir an, es sei $s + 1 \leq p \leq n$; (5.5.8) gelte für alle i mit $s \leq i \leq p - 1$. Wir benutzen

$$e_p^t \tilde{L} = \sum_{j=s}^{p} \tilde{l}_{p,j} e_j^t\,,$$

wobei $\tilde{l}_{p,p} = 1$ ist; daher haben wir

$$e_p^t \tilde{M}_k = \sum_{j=s}^{p} \tilde{l}_{p,j} e_j^t F_k = e_p^t F_k + \sum_{j=s}^{p-1} \tilde{l}_{p,j} e_j^t F_k\,,$$

mithin

$$e_p^t F_k = e_p^t \tilde{M}_k - \sum_{j=s}^{p-1} \tilde{l}_{p,j} (e_j^t F_k)\,.$$

Hieraus gewinnen wir (5.5.8) für $i = p$, wenn wir die Abschätzung der $\tilde{M}_k$ beachten und die Induktionsannahme benutzen.

Mit (5.5.8), speziell auf $i = n$ angewendet, ist der Beweis des Hilfssatzes abgeschlossen.

(5.5.9) **Hilfssatz.** *Es seien für* $k \in \mathbb{N}$ $G_k \in M(n \times n, \mathbb{C})$ *mit*

$$I + G_k \text{ invertierbar}, \quad \lim_{k \to \infty} G_k = 0 .$$

Dann existieren QR*-Zerlegungen*

$$I + G_k = \tilde{Q}_k \tilde{R}_k \quad (k \in \mathbb{N})$$

mit der Eigenschaft

$$\lim_{k \to \infty} \tilde{Q}_k = I , \quad \lim_{k \to \infty} \tilde{R}_k = I .$$

Hat man darüber hinaus $|G_k| = O(\rho^k)$ $(k \to \infty)$, *so gilt*

$$|\tilde{Q}_k - I| = O(\rho^k) \quad (k \to \infty) .$$

Im Fall $|e_n^t G_k| = O(\rho_n^k)$ $(k \to \infty)$ *gewinnen wir mit* $\alpha_k \in \mathbb{C}, \neq 0$ *sowie* $v_k \in \mathbb{C}^n$ *eine Darstellung*

$$(5.5.10) \quad \begin{cases} e_n^t \tilde{Q}_k = \alpha_k e_n^t + v_k^t \quad (k \in \mathbb{N}) , \\ \lim_{k \to \infty} \alpha_k = 1 , \ |v_k| = O(\rho_n^k) \quad (k \to \infty) . \end{cases}$$

Beweis. Ausgehend von QR-Zerlegungen

$$I + G_k = Q_k R_k , \ R_k = (r_{i,j}^{(k)})_{(n,n)} \text{ invertierbar} \quad (k \in \mathbb{N}) ,$$

definieren wir die unitären Diagonalmatrizen

$$D_k := \operatorname{diag}\left(\frac{\bar{r}_{1,1}^{(k)}}{|r_{1,1}^{(k)}|} , \ldots , \frac{\bar{r}_{n,n}^{(k)}}{|r_{n,n}^{(k)}|} \right)$$

und setzen

$$\tilde{Q}_k := Q_k D_k^{-1} , \ \tilde{R}_k = (\tilde{r}_{i,j}^{(k)})_{(n,n)} := D_k R_k .$$

Somit gewinnen wir QR-Zerlegungen $I + G_k = \tilde{Q}_k \tilde{R}_k$ $(k \in \mathbb{N})$ mit der zusätzlichen Eigenschaft

$$\tilde{r}_{i,i}^{(k)} > 0 \qquad (i = 1, \ldots, n; \ k \in \mathbb{N}) .$$

Wir berechnen nun

$$(I + G_k)^{-1} = I - G_k (I + G_k)^{-1} =: I + H_k \quad (k \in \mathbb{N}) ;$$

nach Satz (3.2.13) bildet $|(I + G_k)^{-1}|$ $(k \in \mathbb{N})$ eine beschränkte Folge, aufgrunddessen

$$\lim_{k \to \infty} H_k = 0$$

bzw. sogar $|H_k| = O(\rho^k)$ $(k \to \infty)$ gilt. Aus den Beziehungen

$$\tilde{Q}_k \tilde{R}_k = I + G_k\,, \quad \tilde{R}_k^{-1} \tilde{Q}_k^* = I + H_k$$

folgern wir – vgl. (3.7.21) –

$$|\tilde{R}_k| = |I + G_k| \leqq \sqrt{n} + |G_k|\,, \; |\tilde{R}_k^{-1}| = |I + H_k| \leqq \sqrt{n} + |H_k| \quad (k \in \mathbb{N})\,;$$

mithin sind die Folgen der $|\tilde{R}_k|$ und $|\tilde{R}_k^{-1}|$ beschränkt. Außerdem ergibt sich

$$\tilde{Q}_k = \tilde{R}_k^{-1} + G_k \tilde{R}_k^{-1} \tag{5.5.11}$$

bzw. $\tilde{Q}_k^* = \tilde{R}_k + \tilde{R}_k H_k$, also

$$\tilde{Q}_k = \tilde{R}_k^* + H_k^* \tilde{R}_k^*\,. \tag{5.5.12}$$

Hierin haben wir

$$\lim_{k \to \infty} (G_k \tilde{R}_k^{-1}) = \lim_{k \to \infty} (H_k^* \tilde{R}_k^*) = 0$$

und im Fall $|G_k| = O(\rho^k)$ schärfer

$$|G_k \tilde{R}_k^{-1}| = O(\rho^k)\,, \; |H_k^* \tilde{R}_k^*| = O(\rho^k) \quad (k \to \infty)\,.$$

Weiter wollen wir

$$\left\{ \begin{array}{l} \lim\limits_{k \to \infty} \tilde{R}_k = I\,, \\ |\tilde{R}_k - I| = O(\rho^k) \text{ im Fall } |G_k| = O(\rho^k) \quad (k \to \infty) \end{array} \right. \tag{5.5.13}$$

zeigen; dann folgen wegen (5.5.12) die entsprechenden Aussagen auch für $\tilde{Q}_k$.

Aus (5.5.11) zusammen mit (5.5.12) gewinnen wir

$$\tilde{R}_k^* = \tilde{R}_k^{-1} + G_k \tilde{R}_k^{-1} - H_k^* \tilde{R}_k^* =: \tilde{R}_k^{-1} + U_k \quad (k \in \mathbb{N}) \tag{5.5.14}$$

mit $\lim\limits_{k \to \infty} U_k = 0$ bzw. $|U_k| = O(\rho^k)$ $(k \to \infty)$. Da die $\tilde{R}_k^*$ untere, die $\tilde{R}_k^{-1}$ obere Dreiecksmatrizen sind, ergibt sich hieraus zunächst für alle $i < j$

$$\tilde{r}_{i,j}^{(k)} \to 0 \quad \text{bzw.} \quad |\tilde{r}_{i,j}^{(k)}| = O(\rho^k) \quad (k \to \infty)\,.$$

Nunmehr definieren wir

$$I + E_k := \operatorname{diag}(\tilde{r}_{1,1}^{(k)}, \ldots, \tilde{r}_{n,n}^{(k)}) \quad (k \in \mathbb{N})\,.$$

Dann sind die $|I + E_k| \leqq |\tilde{R}_k|$, also die Folge der $|I + E_k|$ beschränkt; wegen $\tilde{r}_{i,i}^{(k)} > 0$ gilt $(I + E_k)^* = I + E_k$. Wenn wir in (5.5.14) die Diagonalanteile betrachten, so ergibt sich

$$(I + E_k) = (I + E_k)^{-1} + V_k$$

mit $V_k \to 0$ bzw. $|V_k| = O(\rho^k)$ $(k \to \infty)$. Durch Multiplikation mit $(I + E_k)$ erhalten wir

$$(5.5.15) \qquad (I + E_k)^2 = I + (I + E_k)V_k \,,$$

woraus $\lim_{k \to \infty} (I + E_k)^2 = I$ und wegen $\tilde{r}_{i,i}^{k} > 0$ schärfer

$$\lim_{k \to \infty} (I + E_k) = I$$

folgt. Hiermit ist der 1. Teil von (5.5.13) bewiesen. Aus (5.5.15) gewinnen wir sofort

$$2E_k(I + \tfrac{1}{2}E_k) = (I + E_k)V_k \,,$$

dabei ist wegen $\lim_{k \to \infty} E_k = 0$ $(I + \frac{1}{2}E_k)$ für genügend großes k $(\geqq k_0)$ invertierbar, daher

$$E_k = \tfrac{1}{2}(I + \tfrac{1}{2}E_k)^{-1}(I + E_k)V_k \qquad (k \geqq k_0) \,.$$

Aus $|G_k| = O(\rho^k)$ folgt, wie schon bemerkt, $|V_k| = O(\rho^k)$ $(k \to \infty)$, außerdem bleibt die Folge der $|(I + \frac{1}{2}E_k)^{-1}(I + E_k)|$ $(k \geqq k_0)$ beschränkt; daher haben wir

$$|E_k| = O(\rho^k) \qquad (k \to \infty)$$

womit auch die zweite Aussage von (5.5.13) gezeigt ist.

Nun zu (5.5.10): Nach (5.5.11) haben wir

$$e_n^t \tilde{Q}_k = e_n^t \tilde{R}_k^{-1} + e_n^t G_k \tilde{R}_k^{-1} = \frac{1}{\tilde{r}_{n,n}^{(k)}} e_n^t + (e_n^t G_k)\tilde{R}_k^{-1} \,.$$

Dabei strebt $\alpha_k := \dfrac{1}{\tilde{r}_{n,n}^{(k)}} \to 1$ für $k \to \infty$; schließlich genügen die $v_k^t := (e_n^t G_k)\tilde{R}_k^{-1}$ der Abschätzung

$$|v_k^t| \leqq |e_n^t G_k| \cdot |\tilde{R}_k^{-1}| = O(\rho_n^k) \qquad (k \to \infty) \,,$$

letzteres, da die $|\tilde{R}_k^{-1}|$ beschränkt sind.

Nach diesen Vorbereitungen beweisen wir zur Konvergenz des QR-Verfahrens den

(5.5.16) **Satz.** *Zu vorgegebenem* $A \in M(n \times n, \mathbb{C})$ *seien für* $k \in \mathbb{N}$ *die Matrizen* $A_k = (a_{i,j}^{(k)})_{(n,n)}$ *gemäß* (5.5.1) *konstruiert;* A *erfülle die Voraussetzung* (5.5.3).

Dann behaupten wir

(i) *Es existieren unitäre Diagonalmatrizen*

$$S_k = \operatorname{diag}(\sigma_1^{(k)}, \ldots, \sigma_n^{(k)}) \quad (\sigma_i^{(k)} \in \mathbb{C},\ |\sigma_i^{(k)}| = 1) \quad (k \in \mathbb{N}),$$

so daß

$$\lim_{k \to \infty} S_k A_k S_k^* = \Lambda$$

gilt, worin Λ *eine obere Dreiecksmatrix mit den Diagonalelementen* $\lambda_1, \ldots, \lambda_n$ *ist. Insbesondere haben wir*

$$\lim_{k \to \infty} a_{i,i}^{(k)} = \lambda_i \qquad (i = 1, \ldots, n).$$

(ii) *Die Konvergenzgeschwindigkeit beträgt*

$$\left.\begin{aligned} |S_k A_k S_k^* - \Lambda| &= O(\rho^k) \\ |e_n^t A_k - \lambda_n e_n^t| &= O(\rho_n^k) \end{aligned}\right\} \quad (k \to \infty).$$

Beweis. Aus $D = T^{-1}AT$ folgern wir induktiv

$$A^k = TD^kT^{-1} \quad (k \in \mathbb{N}).$$

Die Zerlegung von T^{-1} liefert

$$A^k = TD^kLU = T(D^kLD^{-k})D^kU$$

und Hilfssatz (5.5.5) weiter

$$A^k = T\tilde{L}(I + F_k)D^kU.$$

Wir verwenden nun eine Zerlegung

$$T\tilde{L} = QR$$

mit unitärer Matrix Q und oberer Dreiecksmatrix R; hiermit wird

$$A^k = QR(I + F_k)D^kU = Q(I + RF_kR^{-1})RD^kU. \tag{5.5.17}$$

Wir setzen

$$G_k := RF_kR^{-1} \quad (k \in \mathbb{N}) \tag{5.5.18}$$

und rechnen auf Grund der entsprechenden Eigenschaft von F_k sofort

$$|G_k| = O(\rho^k) \quad (k \to \infty)$$

nach. Außerdem haben wir

$$e_n^tG_k = (e_n^tR)F_kR^{-1} = r_{n,n}(e_n^tF_k)R^{-1},$$

woraus unmittelbar

$$|e_n^tG_k| = O(\rho_n^k) \quad (k \to \infty)$$

folgt. Schließlich sind die Matrizen $I + G_k$ invertierbar, da nach Konstruktion die $I + F_k$ diese Eigenschaft haben. Wir können daher den Hilfssatz (5.5.9) anwenden und erhalten demgemäß die dort näher beschriebenen Zerlegungen

$$I + G_k = \tilde{Q}_k \tilde{R}_k .$$

Einsetzen in (5.5.17) liefert unter Berücksichtigung von (5.5.18)

(5.5.19) $\quad A^k = (Q\tilde{Q}_k)(\tilde{R}_k R D^k U) .$

Offenbar definieren (5.5.2), (ii) und (5.5.19) je eine QR-Zerlegung von A^k. Nach dem Eindeutigkeitssatz (2.6.23) existiert für jedes $k \in \mathbb{N}$ eine unitäre Diagonalmatrix S_k mit

(5.5.20) $\quad (Q\tilde{Q}_k) S_k = (Q_0 \cdots Q_{k-1}) .$

Mit (5.5.2), (i) führt dies zu

(5.5.21) $\quad A_k = S_k^* \tilde{Q}_k^* Q^* A \, Q \, \tilde{Q}_k S_k .$

Aus der Voraussetzung (5.5.3), (i) leiten wir unter Beachtung der Beziehungen $T = QR\tilde{L}^{-1}$ und $D\tilde{L} = \tilde{L}D$ die Darstellung

$$A = TDT^{-1} = QR\tilde{L}^{-1} D\tilde{L} R^{-1} Q^* = QRDR^{-1} Q^*$$

ab. Eingesetzt in (5.5.21), ergibt sich so

(5.5.22) $\quad S_k A_k S_k^* = \tilde{Q}_k^* (RDR^{-1}) \tilde{Q}_k .$

Da nach Konstruktion

$$\lim_{k \to \infty} \tilde{Q}_k = I$$

gilt, streben die $S_k A_k S_k^*$ für $k \to \infty$ gegen die obere Dreiecksmatrix

$$\Lambda := RDR^{-1} ,$$

die offenbar die Diagonalelemente $\lambda_1, \ldots, \lambda_n$ besitzt. Dabei gilt

$$|S_k A_k S_k^* - \Lambda| = |\tilde{Q}_k^* \Lambda \tilde{Q}_k - \Lambda| \leqq |\tilde{Q}_k^* \Lambda \tilde{Q}_k - \tilde{Q}_k^* \Lambda| + |\tilde{Q}_k^* \Lambda - \Lambda|$$
$$\leqq |\Lambda| \, |\tilde{Q}_k - I| + |\tilde{Q}_k^* - I| \, |\Lambda| = O(\rho^k) \quad (k \to \infty) .$$

Es bleibt das Konvergenzverhalten der n-ten Zeile zu untersuchen: Nach (5.5.22) erhalten wir

$$e_n^t \tilde{Q}_k (S_k A_k S_k^*) = e_n^t \Lambda \tilde{Q}_k = \lambda_n e_n^t \tilde{Q}_k$$

und weiter gemäß (5.5.10)

$$\alpha_k e_n^t (S_k A_k S_k^*) + v_k^t (S_k A_k S_k^*) = \alpha_k \lambda_n e_n^t + \lambda_n v_k^t ,$$

mithin

$$e_n^t (S_k A_k S_k^*) - \lambda_n e_n^t = \frac{1}{\alpha_k} v_k^t (\lambda_n I - S_k A_k S_k^*) .$$

Es folgt, wie man durch komponentenweises Rechnen feststellt,

$$|e_n^t A_k - \lambda_n e_n^t| = |e_n^t S_k A_k S_k^* - \lambda_n e_n^t|$$

$$\leqq \frac{1}{|\alpha_k|}(|\lambda_n| + |A_k|)\,|v_k^t| = O(\rho_n^k) \quad (k \to \infty),$$

letzteres, da die Folge der $|\alpha_k|^{-1}$ beschränkt ist und nach (5.5.2), (i) $|A_k| = |A|$ gilt.

Wenn die Bedingung (5.5.3), (i) verletzt ist, also die Jordansche Normalform zu A Einsen in der Nebendiagonalen besitzt, so bleibt das QR-Verfahren – i.a. erheblich langsamer – konvergent; hierzu vgl. Wilkinson [61], S. 541 ff.

Wenn andererseits T^{-1} keine Dreieckszerlegung besitzt, aber beispielsweise die Voraussetzungen (5.5.3), (i) und (iii) gegeben sind, so definiert man eine Permutationsmatrix P, mit der $PT^{-1} = LU$ gilt, durch folgendes Vorgehen: man wendet das Gaußsche Eliminationsverfahren auf $C := T^{-1}$ an, wobei man – ähnlich wie in (2.2.8) – als Pivotelement beim ν-ten Eliminationsschritt jeweils $c_{i_\nu,\nu}^{(\nu)}$ mit

$$i_\nu = \min\{i \in \{\nu, \ldots, n\} : c_{i,\nu}^{(\nu)} \neq 0\}$$

wählt. Dann besitzen mit $D' = \operatorname{diag}(\lambda_1', \ldots, \lambda_n') := PDP^{-1}$ die Matrizen $D'^k L D'^{-k}$ ein ähnliches Konvergenzverhalten wie in Hilfssatz (5.5.5), da nach Konstruktion für alle $i \geqq j$ mit $|\lambda_i'| > |\lambda_j'|$ die Koeffizienten $l_{i,j}$ verschwinden. Der Satz (5.5.16) bleibt also im Kern gültig; es ändert sich lediglich die Reihenfolge der λ_i in der Diagonalen von Λ. Dies ist bei Wilkinson [61], S. 519 diskutiert.

Ausführlicher gehen wir auf eine Modifikation der Voraussetzung (5.5.3), (iii) ein; wir fordern nämlich die schwächere Bedingung

(iii′) Es existieren keine Indizes $i_1, i_2, i_3 \in \{1, \ldots, n\}$ mit

$$\begin{cases} \lambda_{i_1}, \lambda_{i_2}, \lambda_{i_3} & \text{paarweise verschieden,} \\ |\lambda_{i_1}| = |\lambda_{i_2}| = |\lambda_{i_3}| . \end{cases}$$

Die Voraussetzung (iii′) wird – im Gegensatz zu (iii) – eventuell auch von einer reellen, nichtsymmetrischen Matrix A mit Paaren konjugiert-komplexer Eigenwerte erfüllt. Ferner ist (iii′) stets dann gegeben, wenn A nur reelle Eigenwerte besitzt, beispielsweise hermitesch ist.

Der Einfachheit halber setzen wir außer (5.5.3), (i), (ii) den folgenden Spezialfall von (iii′) voraus:

(iii*) Es existiere ein $r \in \{1, 2, \ldots, n-1\}$ mit

$$\begin{cases} |\lambda_{r-1}| > |\lambda_r| = |\lambda_{r+1}| > |\lambda_{r+2}| , \\ \lambda_r \neq \lambda_{r+1} ; \end{cases}$$

ferner gelte für alle $i \in \{1, \ldots, n-1\}$ mit $i \neq r$

$$|\lambda_i| = |\lambda_{i+1}| \Rightarrow \lambda_i = \lambda_{i+1} .$$

Hierbei sei formal $|\lambda_0| = +\infty$, $|\lambda_{n+1}| = -\infty$ gesetzt. Da wir in (5.5.3), (i) $|\lambda_1| > |\lambda_n|$ fordern, können wir die Größen ρ und ρ_n aus (5.5.4) übernehmen. Außerdem werden wir die Matrizen M_k und $\tilde{L}$ aus dem Beweis des Hilfssatzes (5.5.5) verwenden. Weiter beachten wir, daß hier mit einem $\alpha \in \mathbb{C}$ und einem $\varphi \in \mathbb{R}$ für $k \in \mathbb{N}$

$$m_{r+1,r}^{(k)} = l_{r+1,r}\left(\frac{\lambda_{r+1}}{\lambda_r}\right)^k = \alpha e^{ik\varphi} =: \omega_k ,$$

mithin $|\omega_k| = |\alpha|$ gilt. Entsprechend setzen wir

(5.5.23) $\quad J_k := I + \omega_k e_{r+1} e_r^t .$

In Analogie zu Hilfssatz (5.5.5) beweisen wir die

(5.5.24) **Bemerkung.** Es gilt

$$D^k L D^{-k} = \tilde{L} J_k (I + F_k) \quad (k \in \mathbb{N}) ,$$

wobei die Matrizen F_k für $k \to \infty$ das Konvergenzverhalten

$$|F_k| = O(\rho^k), \quad |e_n^t F_k| = O(\rho_n^k)$$

besitzen.

Beweis. Definieren wir $\tilde{M}_k$ durch

$$M_k = \tilde{L} + (J_k - I) + \tilde{M}_k ,$$

so besitzen diese $\tilde{M}_k$ das gleiche Konvergenzverhalten wie die ebenso bezeichneten Matrizen im Beweis zu (5.5.5). Wegen $\lambda_i \neq \lambda_{r+1}$ $(i \neq r+1)$ gilt $\tilde{L} e_{r+1} = e_{r+1}$, mithin

$$\tilde{L} J_k = \tilde{L} + \omega_k (\tilde{L} e_{r+1}) e_r^t = \tilde{L} + (J_k - I)$$

und infolgedessen mit $F_k := (\tilde{L} J_k)^{-1} \tilde{M}_k$

$$M_k = \tilde{L} J_k (I + F_k) .$$

Da $J_k^{-1} = I - \omega_k e_{r+1} e_r^t$, also

$$|J_k^{-1}| \leqq (n + |\alpha|^2)^{\frac{1}{2}} \quad (k \in \mathbb{N})$$

und danach die Folge der $|(\tilde{L} J_k)^{-1}| = |J_k^{-1} \tilde{L}^{-1}|$ beschränkt ist, erhält man – wie behauptet – $|F_k| = O(\rho^k)$ $(k \to \infty)$.

Es bleibt die Untersuchung der n-ten Zeile der F_k. Dazu sei s gemäß (5.5.7) bestimmt.

Wir betrachten zunächst den Fall $r + 1 < n$: Hier ergibt sich wegen $|\lambda_{r+1}| > |\lambda_{r+2}|$ sicherlich $r + 1 < s$ und folglich

$$e_i^t \tilde{L} J_k = \sum_{j=s}^{i} \tilde{l}_{i,j} (e_j^t J_k) = \sum_{j=s}^{i} \tilde{l}_{i,j} e_j^t = e_i^t \tilde{L} \quad (s \leqq i \leqq n) .$$

Mit Hilfe dieser Gleichungen gewinnt man Abschätzungen des Typs (5.5.8) ähnlich wie im Beweis zu (5.5.5).

Im Fall $r+1=n$ haben wir

$$|\lambda_{n-2}| = |\lambda_{r-1}| > |\lambda_r| = |\lambda_{n-1}| = |\lambda_{r+1}| = |\lambda_n|$$

und daher $s = n-1$. Auf Grund der Definition der Matrizen $\tilde{L}, J_k$ folgt der Reihe nach

$$e_{n-1}^t \tilde{L} = e_{n-1}^t, \quad e_n^t \tilde{L} = e_n^t$$

sowie

$$\begin{cases} e_{n-1}^t(\tilde{L}J_k) = e_{n-1}^t J_k = e_{n-1}^t \\ e_n^t(\tilde{L}J_k) = e_n^t J_k = e_n^t + \omega_k e_{n-1}^t \,. \end{cases}$$

Wegen $\tilde{M}_k = (\tilde{L}J_k)F_k$ gewinnen wir hieraus die Beziehungen

$$e_{n-1}^t \tilde{M}_k = e_{n-1}^t F_k, \quad e_n^t \tilde{M}_k = (e_n^t + \omega_k e_{n-1}^t) F_k \,,$$

mithin schließlich

$$e_n^t F_k = e_n^t \tilde{M}_k - \omega_k e_{n-1}^t \tilde{M}_k \qquad (k \in \mathbb{N}) \,.$$

Hierin gilt für $k \to \infty$ neben $e_n^t \tilde{M}_k = O(\rho_n^k)$ wegen $|\lambda_n| = |\lambda_{n-1}|$ auch $e_{n-1}^t \tilde{M}_k = O(\rho_n^k)$, womit der Beweis abgeschlossen ist.

Weiter benötigen wir den

(5.5.25) **Hilfssatz.** *Es seien* $C, J \in M(n \times n, \mathbb{C})$ *und dabei* J *eine Matrix des Typs* (5.5.23), *d.h. mit einem* $\omega \in \mathbb{C}$ *und einem* $r \in \{1, \ldots, n-1\}$

$$J = I + \omega e_{r+1} e_r^t \,;$$

ferner sei eine QR*-Zerlegung*

$$C = QR$$

mit unitärer Matrix Q *und oberer Dreiecksmatrix* R *vorgegeben.*

Dann existiert eine QR*-Zerlegung*

$$CJ = (QV)U \,,$$

in der U *obere Dreiecksmatrix ist und* V *mit einer unitären* (2,2)*-Matrix* W *die Gestalt*

$$(5.5.26) \qquad V = \left(\begin{array}{c|c|c} I_{r-1} & 0 & 0 \\ \hline 0 & W & 0 \\ \hline 0 & 0 & I_{n-r-1} \end{array}\right)$$

besitzt.

Beweis. Laut Voraussetzung haben wir

$$CJ = Q(RJ)$$

mit

$$RJ = R(I + \omega e_{r+1} e_r^t) = R + \omega R e_{r+1} e_r^t .$$

In der Matrix $Re_{r+1}e_r^t$ verschwinden in jedem Fall außer der r-ten alle übrigen Spalten, in der r-ten Spalte steht die (r + 1)-te Spalte der Matrix R. Demgemäß besitzt RJ die Form

$$RJ = \begin{pmatrix} 0 & & \\ & \tilde{u}_{r,r} & \tilde{u}_{r,r+1} \\ 0 & & \\ & \tilde{u}_{r+1,r} & \tilde{u}_{r+1,r+1} \\ 0 & 0 & 0 \end{pmatrix} .$$

Weiter zerlegen wir

$$\begin{pmatrix} \tilde{u}_{r,r} & \tilde{u}_{r,r+1} \\ \tilde{u}_{r+1,r} & \tilde{u}_{r+1,r+1} \end{pmatrix} = WT ,$$

worin W unitär, T obere Dreiecksmatrix ist. Wenn wir zu W die unitäre Matrix V durch (5.5.26) definieren, so wird

$$V^*RJ = \begin{pmatrix} 0 & & \\ 0 & T & \\ 0 & 0 & 0 \end{pmatrix} ,$$

also eine obere Dreiecksmatrix, die wir mit U bezeichnen. Die behauptete Zerlegung von CJ ergibt sich aus den Gleichungen

$$CJ = Q(RJ) = (QV)(V^*RJ) = (QV)U .$$

In Analogie zum Satz (5.5.16) notieren wir nun den

(5.5.27) **Satz.** *Zu vorgegebenem* $A \in M(n \times n, \mathbb{C})$ *seien für* $k \in \mathbb{N}$ *die* $A_k = (a_{i,j}^{(k)})_{(n,n)}$ *gemäß* (5.5.1) *konstruiert. A erfülle die Voraussetzungen* (5.5.3), (i), (ii) *und* (iii*).

Dann behaupten wir:

(i) *Es existieren unitäre Diagonalmatrizen* S_k *und unitäre Matrizen* V_k *der Gestalt* (5.5.26), *so daß*

$$\lim_{k\to\infty}(V_kS_k)\,A_k(V_kS_k)^* = \Lambda$$

gilt, worin Λ *eine obere Dreiecksmatrix mit den Diagonalelementen* $\lambda_1,\dots,\lambda_n$ *ist. Insbesondere haben wir*

$$\lim_{k\to\infty} a_{i,i}^{(k)} = \lambda_i \qquad (i \neq r, r+1)$$

sowie für die Teilmatrizen

$$A_k^{(r)} := \begin{pmatrix} a_{r,r}^{(k)} & a_{r,r+1}^{(k)} \\ a_{r+1,r}^{(k)} & a_{r+1,r+1}^{(k)} \end{pmatrix}$$

von A_k

$$\lim_{k\to\infty}(\det A_k^{(r)}) = \lambda_r\lambda_{r+1};\quad \lim_{k\to\infty}(\operatorname{spur} A_k^{(r)}) = \lambda_r + \lambda_{r+1}\,.$$

(ii) *Die Konvergenzgüte beträgt*

$$\left.\begin{aligned} &|(V_kS_k)A_k(V_kS_k)^* - \Lambda| \\ &|\det A_k^{(r)} - \lambda_r\lambda_{r+1}| \\ &|\operatorname{spur} A_k^{(r)} - (\lambda_r + \lambda_{r+1})| \end{aligned}\right\} = O(\rho^k) \quad (k \to \infty)$$

und, falls $r+1 < n$,

$$|e_n^t A_k - \lambda_n e_n^t| = O(\rho_n^k) \quad (k \to \infty)\,.$$

Der *Beweis* verläuft ähnlich wie der zu Satz (5.5.16). Zunächst gilt

(5.5.28) $$A^k = TD^kT^{-1} = T(D^kL\,D^{-k})D^kU = (T\tilde{L})\,J_k(I + F_k)D^kU \quad (k \in \mathbb{N})\,,$$

letzteres nach (5.5.24). Für festes $k \in \mathbb{N}$ zerlegen wir gemäß Hilfssatz (5.5.25)

(5.5.29) $$(T\tilde{L})\,J_k = (QV_k)\,U_k\,.$$

Diese Beziehung in (5.5.28) eingesetzt, ergibt mit $G_k := U_kF_kU_k^{-1}$

(5.5.30) $$A^k = (QV_k)\,(I + G_k)\,U_kD^kU\,.$$

Mit der (von k unabhängigen) Matrix $R = Q^*T\tilde{L}$ führt (5.5.29) zu

$$U_k = V_k^*R\,J_k\,,$$

aufgrunddessen

$$|U_k| = |RJ_k| \leqq |R|\,(1 + |\alpha|),\ |U_k^{-1}| = |J_k^{-1}R^{-1}| \leqq |R^{-1}|\,(1 + |\alpha|)$$

abschätzbar ist. Da die $I + G_k$ schließlich auch invertierbar sind, besitzen die G_k sämtliche in (5.5.9) geforderten Eigenschaften; demgemäß zerlegen wir

$$I + G_k = \tilde{Q}_k \tilde{R}_k$$

und erhalten nach (5.5.30)

$$A^k = (QV_k\tilde{Q}_k)(\tilde{R}_k U_k D^k U) .$$

Ein Vergleich mit der Darstellung (5.5.2), (ii) liefert wie im Beweis zu (5.5.16) die Existenz unitärer Diagonalmatrizen S_k, so daß mit $\Lambda := RDR^{-1}$ für $k \in \mathbb{N}$

$$S_k A_k S_k^* = \tilde{Q}_k^* V_k^* \Lambda V_k \tilde{Q}_k$$

gilt. Aus $|\tilde{Q}_k - I| = O(\rho^k)$ folgt zunächst, wie gehabt,

$$|S_k A_k S_k^* - V_k^* \Lambda V_k| = O(\rho^k) \quad (k \to \infty)$$

und dann, da die V_k unitär sind, auch

$$(5.5.31) \qquad |(V_k S_k) A_k (V_k S_k)^* - \Lambda| = O(\rho^k) \quad (k \to \infty) .$$

Ebenfalls wie im Beweis zu (5.5.16) erschließen wir die Behauptung bezüglich der Konvergenzgüte der n-ten Zeile. Hierzu ist zu beachten, daß man im Falle $r + 1 < n$ $e_n^t V_k^* = e_n^t V_k = e_n^t$, folglich

$$e_n^t \tilde{Q}_k S_k A_k S_k^* = e_n^t (V_k^* \Lambda V_k) \tilde{Q}_k = \lambda_n e_n^t \tilde{Q}_k$$

hat.

Es bleibt das Konvergenzverhalten der Teilmatrizen $A_k^{(r)}$ zu diskutieren. Dazu betrachten wir diejenigen (2,2)-Teilmatrizen von $(V_k S_k) A_k (V_k S_k)^*$, die an der gleichen Stelle wie die $A_k^{(r)}$ stehen. Diese Matrizen lassen sich mit unitären (2,2)-Matrizen $\tilde{W}_k$ als $\tilde{W}_k A_k^{(r)} \tilde{W}_k^*$ schreiben. Bezeichnet man noch mit

$$\Lambda^{(r)} = \begin{pmatrix} \lambda_r & \lambda_{r,r+1} \\ 0 & \lambda_{r+1} \end{pmatrix}$$

die $A_k^{(r)}$ entsprechende Teilmatrix von Λ, so folgt aus (5.5.31)

$$|\tilde{W}_k A_k^{(r)} \tilde{W}_k^* - \Lambda^{(r)}| = O(\rho^k) \quad (k \to \infty) .$$

Hieraus leiten wir für $k \to \infty$ die asymptotischen Aussagen

$$(5.5.32) \quad \begin{cases} |\operatorname{spur} A_k^{(r)} - (\lambda_r + \lambda_{r+1})| = |\operatorname{spur}(\tilde{W}_k A_k^{(r)} \tilde{W}_k^* - \Lambda^{(r)})| = O(\rho^k) , \\ |\det A_k^{(r)} - \lambda_r \lambda_{r+1}| = |\det(\tilde{W}_k A_k^{(r)} \tilde{W}_k^*) - \det \Lambda^{(r)}| = O(\rho^k) \end{cases}$$

ab, womit der Satz (5.5.27) bewiesen ist.

Ergänzend stellen wir fest, daß wir als Näherungen $\lambda_r^{(k)}$ und $\lambda_{r+1}^{(k)}$ für λ_r bzw. λ_{r+1} die Nullstellen des quadratischen Polynoms

$$\det(\lambda I_2 - A_k^{(r)})$$

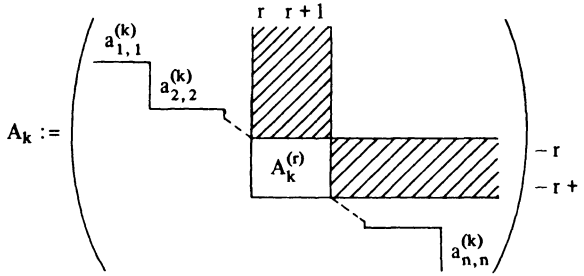
erhalten. Die Koeffizienten dieser Polynome streben nach (5.5.32) für $k \to \infty$ wie ρ^k gegen die Koeffizienten des Polynoms

$$\det(\lambda I_2 - \Lambda^{(r)}) .$$

Da dieses Polynom einfache Nullstellen – nämlich λ_r, λ_{r+1} – besitzt, erschließt man hieraus für $i = r, r+1$

$$|\lambda_i^{(k)} - \lambda_i| = O(\rho^k) \quad (k \to \infty) .$$

Das Konvergenzverhalten der A_k wollen wir an Hand einer Skizze veranschaulichen. Es sei



Wir beobachten:

a) Konvergenz der Koeffizienten unterhalb der Diagonalen mit Ausnahme von $a_{r+1,r}^{(k)}$ gegen Null,

b) Konvergenz der $a_{i,i}^{(k)}$ gegen λ_i für $i \neq r, r+1$,

c) Konvergenz der Eigenwerte von $A_k^{(r)}$ gegen λ_r bzw. λ_{r+1},

d) unregelmäßiges Verhalten (aber Beschränktheit) der Koeffizienten im schraffierten Teil,

e) Konvergenz der Beträge der Koeffizienten im verbleibenden Teil.

Die vorstehenden Überlegungen sind leicht auf die Fälle zu verallgemeinern, in denen die Voraussetzung (iii*) oder sogar (iii') verletzt ist. Es treten dann in jedem A_k an Stelle einer (2,2)-Teilmatrix $A_k^{(r)}$ eventuell mehrere bzw. größere Blöcke auf, deren Eigenwerte für $k \to \infty$ gegen die entsprechenden (betragsgleichen) Eigenwerte von A konvergieren.

Wir geben nun Hinweise zur praktischen Durchführung:

(I) Da die QR-Zerlegungen beliebiger Matrizen einen hohen Rechenaufwand erfordern, wird man eine vorgegebene Matrix zunächst auf obere Hessenberg- bzw.

Tridiagonalform bringen und den Algorithmus (5.5.1) auf die transformierte Matrix anwenden. Dazu überlegt man sich (als Übungsaufgabe 5.6) die

(5.5.33) **Bemerkung.** Ist A obere Hessenbergmatrix (hermitesche Tridiagonalmatrix), so sind alle A_k obere Hessenbergmatrizen (hermitesche Tridiagonalmatrizen).

Im reellen Fall können die QR-Zerlegungen der Hessenbergmatrizen nach der Methode von Givens (Übungsaufgabe 2.11) durchgeführt werden, da hierbei die gleiche Zahl von Rechenoperationen wie beim Householder-Verfahren benötigt wird.

(II) Um die Konvergenz zu beschleunigen, wird – ähnlich wie in 5.1 bei der inversen Potenzmethode – die Technik der Spektralverschiebung benutzt. Ist μ Näherung eines Eigenwerts λ von A, so wendet man das QR-Verfahren auf $A-\mu I$ an.

Ausführlicher betrachten wir den Fall, daß A die Voraussetzungen (5.5.3), (i) und (ii) erfüllt und μ hinreichend nahe beim Eigenwert λ_n von A liegt, so daß für alle $\lambda_i \neq \lambda_n$ die Ungleichungen

$$(5.5.34) \qquad 0 < |\lambda_n - \mu| < |\lambda_i - \mu|$$

gelten. Unter diesen Annahmen konvergieren die n-ten Zeilen der gemäß (5.5.1) aus $A_0(\mu) := A - \mu I$ konstruierten Matrizen $A_k(\mu)$ gegen $(\lambda_n - \mu)e_n^t$. Hierzu überlegen wir uns:

Für $A - \mu I$ ist die Voraussetzung (5.5.3), (i) mit

$$D' = \operatorname{diag}(\lambda_{\pi(1)} - \mu, \ldots, \lambda_{\pi(n)} - \mu) = P^{-1}DP - \mu I\,,$$
$$T' = TP$$

erfüllt, wobei die Permutation $\pi \in S_n$ so gewählt ist, daß

$$|\lambda_{\pi(1)} - \mu| \geq |\lambda_{\pi(2)} - \mu| \geq \ldots \geq |\lambda_{\pi(n)} - \mu| > 0$$

und gemäß (5.5.34)

$$\pi(n) = n$$

gilt; ferner ist die Permutationsmatrix P durch

$$P e_i = e_{\pi(i)} \qquad (i = 1, \ldots, n)$$

definiert. Sollte die Matrix

$$T'^{-1} = P^{-1}T^{-1}$$

keine Dreieckszerlegung besitzen, so ist – wie bereits auf S. 42 erläutert – eine Permutationsmatrix P' anzugeben, so daß $P'T'^{-1}$ in Dreiecksmatrizen L', U' zerlegbar ist und gleichzeitig – mit $D'' := P'D'P'^{-1}$ – die Matrizen $D''^{k}L'D''^{-k}$ ein ähnliches Verhalten wie in (5.5.5) bzw. (5.5.24) zeigen. Nach Konstruktion hat diese Permutationsmatrix P', wie man sich auf Grund von $T^{-1} = LU$ sowie $e_n^t P^{-1} = e_n^t$ überlegt, jedenfalls die Eigenschaft $e_n^t P' = e_n^t$.

Falls schließlich die Voraussetzung (5.5.3), (iii) für $A_0(\mu)$ nur in einer abgeschwächten Form, beispielsweise (iii*) zutrifft, so ist wegen (5.5.34) das dort angegebene r kleiner als $n-1$. Daher – und wegen $e_n^t P' = e_n^t$, wie soeben begründet – ist bezüglich der n-ten Zeilen der $A_k(\mu)$ die schärfere Abschätzung aus (5.5.16), (ii) bzw. (5.5.27), (ii) anwendbar; mit

$$\rho_n = \max \left\{ \frac{|\lambda_n - \mu|}{|\lambda_i - \mu|} : \lambda_i \neq \lambda_n \right\}$$

konvergieren demgemäß für $k \to \infty$ die zugehörigen Koeffizienten $a_{n,i}^{(k)}(\mu)$ wie ρ_n^k gegen 0 bzw. gegen $\lambda_n - \mu$. Offensichtlich ist diese Konvergenz umso rascher, je weniger μ von λ_n entfernt ist.

Auf diesen Überlegungen basiert das

(5.5.35) QR-*Verfahren mit Spektralverschiebung.* Wir setzen

$$A_0 := A ;$$

für $k = 0, 1, 2, \ldots$ zerlegen wir

$$A_k - \mu_k I = Q_k R_k$$

und definieren hiermit

$$A_{k+1} := R_k Q_k + \mu_k I .$$

Hierbei werden die μ_k nach verschiedenen Vorschriften gewählt: Beispielsweise bestimmt man zu einem positiven $\epsilon < 1$

$$(5.5.36) \qquad l := \min \left\{ k \geq 1 : \left| \frac{a_{n,n}^{(k)}}{a_{n,n}^{(k-1)}} - 1 \right| \leq \epsilon \right\}$$

und dazu $\mu_k := 0$ für $k = 0, 1, \ldots, l-1$ sowie

$$(5.5.37) \qquad \mu_k := a_{n,n}^{(k)} \qquad (k = l, l+1, \ldots) .$$

Weitgehender anwendbar ist das folgende Vorgehen: es bezeichne $\lambda_n^{(k)}$ denjenigen Eigenwert von

$$A_k^{(n-1)} = \begin{pmatrix} a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} \\ a_{n,n-1}^{(k)} & a_{n,n}^{(k)} \end{pmatrix} ,$$

der näher bei $a_{n,n}^{(k)}$ liegt, sowie $\tilde{\lambda}_n^{(k-1)}$ denjenigen Eigenwert von $A_{k-1}^{(n-1)}$, der ebenfalls näher bei $a_{n,n}^{(k)}$ (bzw. $\lambda_n^{(k)}$) liegt. Hiermit setzt man

$$(5.5.36') \qquad l := \min \left\{ k \geq 1 : \left| \frac{\lambda_n^{(k)}}{\tilde{\lambda}_n^{(k-1)}} - 1 \right| \leq \epsilon \right\}$$

und $\mu_k := 0$ für $k = 0, 1, \ldots, l-1$ sowie

$$(5.5.37') \qquad \mu_k := \lambda_n^{(k)} \qquad (k = l, l+1, \ldots) .$$

Die Existenz des Minimums in (5.5.36) ist dann gewährleistet, wenn A die Voraussetzung (5.5.3), (i), (ii), (iii) bzw. (iii*) mit $r < n-1$ erfüllt. In diesen Fällen ist auch (5.5.36′) sinnvoll, da die $\lambda_n^{(k)}$ und ebenso die $\tilde{\lambda}_n^{(k-1)}$ für $k \to \infty$ gegen λ_n konvergieren. Das Minimum in (5.5.36′) existiert – im Gegensatz zu (5.5.36) – jedoch auch dann, wenn der Fall (5.5.3), (iii*) mit $r = n-1$ vorliegt.

Das Verfahren (5.5.35) konvergiert sicher dann, wenn A die Voraussetzungen (5.5.3), (i), (ii), (iii) bzw. (iii′) erfüllt und ab einem $k_0 \in \mathbb{N}$ mit einem positiven $\eta < 1$

$$(5.5.38) \qquad 0 < |\mu_k - \lambda_n| \leqq \eta\, |\mu_k - \lambda_i| \qquad (k \geqq k_0,\ \lambda_i \neq \lambda_n)$$

zutrifft: unter dieser Annahme lassen sich die im Anschluß an (5.5.34) auf $A - \mu I$ bezogenen Überlegungen unschwer übertragen – vgl. Übungsaufgabe 5.7 –. Zur Erreichung der Bedingung (5.5.38) haben sich in der Praxis die Vorschriften (5.5.36/37) bzw. (5.5.36′/37′) mit $\epsilon = \frac{1}{3}$ als geeignet erwiesen.

Vollständige Konvergenzbeweise sind bisher nur für den Fall bekannt, daß A hermitesch ist: dann ist das Verfahren (5.5.35) nach Wilkinson [61], S. 548 „schließlich" kubisch konvergent, d.h. sofern μ_l in (5.5.37) bzw. (5.5.37′) hinreichend nahe bei λ_n liegt. Ist A sogar reelle symmetrische Tridiagonalmatrix mit $a_{i+1,i} \neq 0$ $(i = 1, \ldots, n-1)$, so ist das Verfahren mit den Auswahlvorschriften (5.5.36′/37′) bei beliebigem $\epsilon > 0$ mindestens quadratisch konvergent, s. Wilkinson [63].

Ist A reelle Matrix und sind λ_{n-1}, λ_n einfache, zueinander konjugiert komplexe Eigenwerte von A, so ist folgende Variante von (5.5.37′) gebräuchlich: man bestimmt l wie in (5.5.36′) und weiter, wenn beispielsweise $l = 2l'$ $(l' \in \mathbb{N})$, also geradzahlig ist,

$$(5.5.37'') \quad \begin{cases} \mu_{2k} := \lambda_n^{(2k)} \\ \mu_{2k+1} := \bar{\mu}_{2k} \end{cases} \qquad (k = l', l'+1, \ldots)\,,$$

man verwendet also abwechselnd konjugiert-komplexe Verschiebungsparameter. Bei diesem Vorgehen ist der Übergang von A_{2k} auf A_{2k+2} in (5.5.35) so darstellbar, daß nur reelle Rechenoperationen auftreten; hierzu vgl. Wilkinson [61], S. 528 ff. Unter geeigneten Voraussetzungen konvergieren dann die (konjugiert komplexen) Eigenwerte von $A_k^{(n-1)}$ gegen λ_n bzw. λ_{n-1}.

(III) Die Berechnung sämtlicher Eigenwerte $\lambda_1, \ldots, \lambda_n$ einer Matrix A geschieht folgendermaßen: Zunächst bringt man die Matrix A (nach der Transformation auf obere Hessenbergform) mittels (5.5.35) auf die Gestalt

$$A_k \approx \left(\begin{array}{c|c} & \times \\ A_k' & \times \\ & \vdots \\ & \times \\ \hline 0 \ldots 0 & \lambda_n \end{array}\right);$$

anschließend berechnet man λ_{n-1}, indem man auf die Teilmatrix A_k' wiederum die Prozedur (5.5.35) anwendet usw. Erfahrungsgemäß werden dabei für jeden weiteren Eigenwert nur noch wenige (i.a. etwa 4) QR-Schritte benötigt.

Abschließend gehen wir noch kurz auf das LR-Verfahren ein. Ist $A \in M(n \times n, \mathbb{C})$ vorgegeben, so lautet der

(5.5.39) *Algorithmus des* LR-*Verfahrens.* Es sei

$$A_0 := A ;$$

für $k = 0, 1, 2, \ldots$ zerlegen wir

$$A_k = L_k R_k ,$$

wobei die L_k normierte untere Dreiecksmatrizen, die R_k obere Dreiecksmatrizen sind, und definieren hiermit

$$A_{k+1} := R_k L_k .$$

Wie wir im 2. Kapitel gesehen haben, sind LR-Zerlegungen nicht immer möglich oder können zu numerischen Instabilitäten führen; daher ist das LR-Verfahren nur bedingt anwendbar. Zur Konvergenz notieren wir den

(5.5.40) **Satz.** *Zu vorgegebenem* $A \in M(n \times n, \mathbb{C})$ *seien für* $k \in \mathbb{N}$ *die Matrizen* A_k *gemäß* (5.5.35) *konstruiert. A erfülle die Voraussetzung* (5.5.3). *Mit* T *aus* (5.5.3) *und* $\tilde{L}$ *aus* (5.5.6) *existiere eine Dreieckszerlegung von* $T\tilde{L}$, *also eine normierte untere Dreiecksmatrix* M *und eine obere Dreiecksmatrix* R, *so daß*

$$T\tilde{L} = MR .$$

Dann behaupten wir

(i) *Es gilt*

$$\lim_{k \to \infty} A_k = \Lambda ,$$

wobei Λ *eine obere Dreiecksmatrix mit den Diagonalelementen* $\lambda_1, \ldots, \lambda_n$ *ist.*

(ii) *Die Konvergenzgeschwindigkeit beträgt*

$$\left.\begin{aligned} |A_k - \Lambda| &= O(\rho^k) , \\ |e_n^t A_k - \lambda_n e_n^t| &= O(\rho_n^k) \end{aligned}\right\} \quad (k \to \infty) .$$

Der Beweis unterscheidet sich kaum von dem zu Satz (5.5.16); die Ausführung sei daher dem Leser überlassen.

In ähnlicher Weise ist der Satz (5.5.27) auf das LR-Verfahren zu übertragen. Zur praktischen Durchführung gelten analoge Überlegungen wie beim QR-Verfahren.

In der angegebenen Form – also mit vorangehender Transformation auf Hessenbergform und in Verbindung mit Spektralverschiebungen gemäß (5.5.35) – haben sich das QR- und das LR-Verfahren als die geeignetsten Methoden erwiesen, um alle Eigenwerte einer Matrix zu bestimmen.

(5.5.41) **Beispiel.** Wir berechnen die Eigenwerte der im Beispiel (5.4.21) angegebenen (4,4)-Matrix A mit Hilfe des QR-Verfahrens. Zunächst transformieren wir A gemäß Satz (5.3.11) in eine symmetrische Tridiagonalmatrix $A^{(0)}$. Bei Benutzung der in (1.1.16) erwähnten REAL*8-Arithmetik, die etwa einer 16-stelligen Dezimalarithmetik entspricht, wird in der numerisch berechneten Matrix $A^{(0)}$ der Koeffizient

$$|a_{4,3}^{(0)}| < 0{,}5 \cdot 10^{-15} .$$

Daher nehmen wir als erste Eigenwertnäherung direkt den Wert

$$\lambda_4 = a_{4,4}^{(0)} = 5{,}000000000000000 .$$

Anschließend wenden wir das QR-Verfahren auf die (3,3)-Teilmatrix

$$A' := (a_{i,j}^{(0)})_{(3,3)}$$

von $A^{(0)}$ an, und zwar – zum Vergleich – einmal ohne Spektralverschiebung, das andere Mal mit Spektralverschiebung gemäß den Vorschriften (5.5.35/36/37). In beiden Fällen iterieren wir so lange, bis

$$(*) \quad |a_{3,2}^{(k)}| < 0{,}5 \cdot 10^{-15}$$

erreicht ist. Dann setzen wir

$$\lambda_3 := a_{3,3}^{k}$$

und berechnen schließlich die verbleibenden Eigenwerte λ_1, λ_2 durch Lösen der quadratischen Gleichung

$$\det(\lambda I_2 - A_k'') = 0 ,$$

wobei

$$A_k'' = (a_{i,j}^{(k)})_{(2,2)}$$

bedeute.

Beim Verfahren ohne Spektralverschiebung ergibt sich (*) nach

$k = 23$

Iterationen, wir erhalten die Eigenwerte

$$\lambda_3 = -0{,}9999999999999936 ,$$
$$\lambda_2 = 14{,}99999999999994 ,$$
$$\lambda_1 = 5{,}000000000000004 .$$

Im zweiten Fall, also beim Verfahren mit Spektralverschiebung, ist (*) bereits nach

$k = 5$

Iterationen erfüllt; die berechneten Eigenwerte sind

$$\lambda_3 = -0{,}9999999999999942 ,$$
$$\lambda_2 = 14{,}99999999999998 ,$$
$$\lambda_1 = 5{,}000000000000007 .$$

5.6. Das Verfahren von Jacobi

Allgemein sei zunächst zu beliebigem $A \in M(n \times n, \mathbb{C})$, $=(a_{i,j})_{(n,n)}$ die Matrix $\tilde{A}$ gemäß der Unterteilung

$$(5.6.1) \qquad A = \operatorname{diag}(a_{1,1}, \ldots, a_{n,n}) + \tilde{A}$$

definiert.

Unter Benutzung dieser Bezeichnung läßt sich die Idee des Jacobi-Verfahrens folgendermaßen einfach beschreiben. Es sei A eine reelle symmetrische (n,n)-Matrix. Zu A konstruiere man schrittweise unitär ähnliche Matrizen A_k mit der Eigenschaft, daß die zugehörigen $\tilde{A}_k$ für $k \to \infty$ gegen die Nullmatrix konvergieren. Diesen infiniten Prozeß breche man ab, sobald $\tilde{A}_k$ (der Norm nach) hinreichend klein ist, und nehme die Diagonalelemente von A_k als Näherungen der Eigenwerte von A.

Als Norm in $M(n \times n, \mathbb{R})$ werden wir in diesem Abschnitt stets die unter (3.3.8) eingeführte Frobeniusnorm verwenden.

Grundlegend ist der

(5.6.2) **Hilfssatz.** *Sind* $A, A', S \in M(n \times n, \mathbb{R})$, *dabei* S *unitär und*

$$\begin{cases} A = (a_{i,j})_{(n,n)} \\ A' = (a'_{i,j})_{(n,n)} := S^* A S , \end{cases}$$

so gilt entsprechend der Bezeichnung (5.6.1)

$$|\tilde{A}'|^2 + \sum_{i=1}^{n} |a'_{i,i}|^2 = |\tilde{A}|^2 + \sum_{i=1}^{n} |a_{i,i}|^2 .$$

Beweis. Mittels Hilfssatz (3.7.21) erschließen wir

$$|A| = |A^*| = |S^* A^*| = |AS| = |S^*(AS)| = |A'| ,$$

womit die behauptete Gleichung bereits bewiesen ist, da gemäß der Definition der Frobeniusnorm ihre linke Seite $|A'|^2$, ihre rechte Seite $|A|^2$ ist.

Als unitäre Transformationen benutzt man die schon im 2. Kapitel – vgl. Übungsaufgabe 2.11 – betrachteten ebenen Drehungen. Analytisch sind diese reell-unitären Matrizen mit (dem Drehwinkel) $\alpha \in \mathbb{R}$ und den Indizes (der Koordinaten der Drehebene) $p, q \in \{1, \ldots, n\}$ mit $p \neq q$ durch

$$(5.6.3) \qquad S(p, q; \alpha) := I + (\cos\alpha - 1)(e_p e_p^t + e_q e_q^t) + \sin\alpha (e_p e_q^t - e_q e_p^t)$$

gegeben. Ihre Transformationseigenschaft läßt sich formelmäßig wie folgt darstellen:

(5.6.4) **Hilfssatz.** *Ist* $A = (a_{i,j})_{(n,n)} \in M(n \times n, \mathbb{R})$, *symmetrisch,*

$$A' = (a'_{i,j})_{(n,n)} := S(p, q; \alpha)^* A\, S(p, q; \alpha) ,$$

so gilt

$$\text{(i)} \quad \begin{cases} a'_{p,j} = a'_{j,p} = a_{p,j}\cos\alpha - a_{q,j}\sin\alpha \\ a'_{q,j} = a'_{j,q} = a_{q,j}\cos\alpha + a_{p,j}\sin\alpha \end{cases} \qquad (j \neq p, q)\,,$$

$$\text{(ii)} \quad \begin{cases} a'_{p,p} = a_{p,p}\cos^2\alpha - 2\,a_{p,q}\cos\alpha\sin\alpha + a_{q,q}\sin^2\alpha\,, \\ a'_{q,q} = a_{q,q}\cos^2\alpha + 2\,a_{p,q}\cos\alpha\sin\alpha + a_{p,p}\sin^2\alpha\,, \end{cases}$$

$$\text{(iii)} \quad a'_{p,q} = a'_{q,p} = (a_{p,p} - a_{q,q})\cos\alpha\sin\alpha + a_{p,q}(\cos^2\alpha - \sin^2\alpha)$$
$$= \tfrac{1}{2}(a_{p,p} - a_{q,q})\sin(2\alpha) + a_{p,q}\cos(2\alpha)\,,$$

$$\text{(iv)} \quad a'_{i,j} = a_{i,j} \qquad (i \neq p, q; j \neq p, q)\,.$$

Der *Beweis* folgt unmittelbar aus der Beziehung

$$a'_{i,j} = e_i^t A' e_j \qquad (i, j = 1, \ldots, n)$$

sowie der analogen für die Matrix A.

Wir wollen nun $\sin\alpha$ und $\cos\alpha$ so bestimmen – α selbst benötigen wir nicht explizit –, daß $|\tilde{A}'|$ möglichst klein wird. Aus (5.6.2) und (5.6.4), (iv) gewinnen wir zunächst

$$|\tilde{A}'|^2 = |\tilde{A}|^2 + \sum_{i=1}^{n}(a_{i,i}^2 - a_{i,i}'^2) = |\tilde{A}|^2 + (a_{p,p}^2 + a_{q,q}^2) - (a_{p,p}'^2 + a_{q,q}'^2)\,.$$

Da offenbar mit einer unitären (2,2)-Matrix $R(\alpha)$ die Gleichung

$$\begin{pmatrix} a'_{p,p} & a'_{p,q} \\ a'_{q,p} & a'_{q,q} \end{pmatrix} = R(\alpha)^* \begin{pmatrix} a_{p,p} & a_{p,q} \\ a_{q,p} & a_{q,q} \end{pmatrix} R(\alpha)$$

gilt, schließen wir aus Hilfssatz (5.6.2)

$$(a_{p,p}'^2 + a_{q,q}'^2) + 2\,a_{p,q}'^2 = (a_{p,p}^2 + a_{q,q}^2) + 2\,a_{p,q}^2$$

und hiermit weiter

$$|\tilde{A}'|^2 = |\tilde{A}|^2 + 2\,a_{p,q}'^2 - 2\,a_{p,q}^2\,. \tag{5.6.5}$$

Demnach erreichen wir den kleinstmöglichen Wert für $|\tilde{A}'|$, wenn

$$a'_{p,q} = 0$$

wird. Dazu setzen wir im Folgenden

$$a_{p,q} \neq 0$$

voraus, da andernfalls $|\tilde{A}'|$ gegenüber $|\tilde{A}|$ nicht echt verkleinert wird.

Zu

$$\vartheta := \frac{a_{q,q} - a_{p,p}}{2\,a_{p,q}}$$

ist nach (5.6.4), (iii) α so zu bestimmen, daß

(5.6.6) $\quad \cot(2\alpha) = \vartheta$

gilt. Dieser Beziehung genügt α wegen

$$\cot(2\alpha) = \frac{1 - \tan^2\alpha}{2\tan\alpha}$$

genau dann, wenn $t = \tan\alpha$ eine Wurzel der quadratischen Gleichung

$$t^2 + 2\,\vartheta\,t - 1 = 0$$

wird. Als t wählen wir die betragsmäßig kleinere Lösung, wozu wir die numerisch stabile Darstellung (1.3.7), (iii), also

$$(5.6.7) \qquad t = \begin{cases} \dfrac{1}{\vartheta + \operatorname{sign}\vartheta\sqrt{1+\vartheta^2}} & (\vartheta \neq 0)\,, \\ 1 & (\vartheta = 0) \end{cases}$$

benutzen. Sollte – bei sehr großem ϑ – die Berechnung von ϑ^2 einen Exponentenüberlauf in der Maschine erzeugen, so ersetzen wir natürlich $\operatorname{sign}\vartheta\sqrt{1+\vartheta^2}$ durch ϑ. Schließlich gewinnen wir

$$(5.6.8) \qquad \begin{cases} \cos\alpha := \dfrac{1}{\sqrt{1+t^2}}\,, \\ \sin\alpha := t\cdot\cos\alpha \end{cases}$$

als mögliche Werte, mit denen $a'_{p,q} = 0$ erreicht wird.

Es ist zweckmäßig, die Gleichungen (5.6.4) – mit $\cos\alpha, \sin\alpha$ gemäß (5.6.8) – in eine numerisch geeignetere Form zu bringen. Hierzu setzen wir

$$u := \frac{\sin\alpha}{1+\cos\alpha}$$

und erhalten

$$(5.6.9) \qquad \begin{cases} \text{(i)} & \begin{cases} a'_{p,j} = a'_{j,p} = a_{p,j} - \sin\alpha\,(a_{q,j} + u\,a_{p,j})\,, \\ a'_{q,j} = a'_{j,q} = a_{q,j} - \sin\alpha\,(a_{p,j} - u\,a_{q,j}) \end{cases} & (j \neq p,q)\,, \\ \text{(ii)} & \begin{cases} a'_{p,p} = a_{p,p} - t\,a_{p,q}\,, \\ a'_{q,q} = a_{q,q} + t\,a_{p,q}\,, \end{cases} & \\ \text{(iii)} & a'_{p,q} = a'_{q,p} = 0\,, & \\ \text{(iv)} & a'_{i,j} = a_{i,j} & (i \neq p,q;\ j \neq p,q)\,. \end{cases}$$

Zur Herleitung von (5.6.9), (i) aus (5.6.4), (i) beachten wir

$$u \sin\alpha = \frac{\sin^2\alpha}{1+\cos\alpha} = 1 - \cos\alpha\,.$$

Zu (5.6.9), (ii) schließen wir aus (5.6.4), (ii) zunächst

$$a'_{p,p} = a_{p,p} + (a_{q,q} - a_{p,p})\sin^2\alpha - 2\,a_{p,q}\cos\alpha\sin\alpha\,;$$

ferner benutzen wir (5.6.4), (iii) mit $a'_{p,q} = 0$, d.h.

$$(a_{q,q} - a_{p,p}) = a_{p,q}\,\frac{\cos^2\alpha - \sin^2\alpha}{\cos\alpha\sin\alpha}\,,$$

womit sich

$$a'_{p,p} = a_{p,p} - a_{p,q}\,\frac{\sin\alpha}{\cos\alpha} = a_{p,p} - t\,a_{q,q}$$

ergibt. Analog beweist man die angegebene Formel für $a'_{q,q}$.

Die Form (5.6.9) erweist sich als vorteilhaft für die numerische Auswertung, da insbesondere für kleine Werte von $|\sin\alpha|$ Fehlerdämpfung – vgl. (1.3.8) – eintritt.

Wir nennen den Übergang

$$A \mapsto A' = S(p,q;\alpha)^* A\, S(p,q;\alpha)$$

gemäß dem Konstruktionsverfahren (5.6.7) bis (5.6.9) einen *Jacobi-Schritt mit dem Pivotelement* $a_{p,q}$. Für die so transformierte Matrix A' gilt nach (5.6.5) natürlich

$$(5.6.10)\qquad |\tilde{A}'|^2 = |\tilde{A}|^2 - 2\,|a_{p,q}|^2\,.$$

Das Jacobi-Verfahren besteht nun aus der sukzessiven Anwendung von Jacobi-Schritten, d.h. aus der Konstruktion

$$(5.6.11)\qquad \begin{cases} A_0 := A \\ A_{k+1} := S_k^* A_k S_k \qquad (k = 0, 1, 2, \ldots)\,, \end{cases}$$

wobei für jedes $k \in \mathbb{N}$ die Transformationsmatrix

$$S_k := S(p_k, q_k; \alpha_k)$$

zu einem Pivotelement $a^{(k)}_{p_k,q_k} \neq 0$ von $A_k = (a^{(k)}_{i,j})_{(n,n)}$ gehört.

Das Verfahren von Jacobi ist im Fall $n \geq 3$ ein infinites Verfahren, da eine in A_{k+1} $(k \in \mathbb{N})$ an der Stelle (p_k, q_k) erzeugte Null bei späteren Transformationsschritten i.a. wieder zerstört wird.

Zur Konvergenz des Verfahrens bei maximaler Pivotwahl beweisen wir den

(5.6.12) **Satz.** *Es sei* $A \in M(n \times n, \mathbb{R})$, *symmetrisch* $(n \geq 3)$; *die* A_k $(k \in \mathbb{N})$ *seien nach* (5.6.11) *konstruiert, wobei für jedes* $k \in \mathbb{N}$ $p_k \neq q_k$ *mit*

$$|a^{(k)}_{p_k, q_k}| = \max_{i \neq j} |a^{(k)}_{i,j}|$$

gewählt sei. Dann haben wir

$$\lim_{k \to \infty} \tilde{A}_k = 0 .$$

Beweis. Für die Frobeniusnorm von $\tilde{A}_k$ $(k \in \mathbb{N})$ gilt offenbar

$$|\tilde{A}_k|^2 \leq n(n-1)\, |a^{(k)}_{p_k, q_k}|^2 .$$

Hieraus ergibt sich wegen (5.6.10) die Abschätzung

$$|\tilde{A}_{k+1}|^2 = |\tilde{A}_k|^2 - 2\, |a^{(k)}_{p_k, q_k}|^2 \leq \left(1 - \frac{2}{n(n-1)}\right) |\tilde{A}_k|^2$$

und demnach mit

$$q := 1 - \frac{2}{n(n-1)} \qquad (0 < q < 1)$$

das Konvergenzverhalten

$$|\tilde{A}_k|^2 \leq q^k\, |\tilde{A}_0|^2 \to 0 \quad (k \to \infty).$$

Bei dem angegebenen Verfahren müssen vor jeder Transformation alle Koeffizienten $a^{(k)}_{i,j}$ mit $i < j$ (A_k symmetrisch!) verglichen werden. Da dies einen erheblichen Zeitaufwand bedeutet, bevorzugt man einfachere Vorschriften für die Pivotwahl.

Beim zyklischen Jacobi-Verfahren werden als Pivotelemente nacheinander die Koeffizienten

$$(5.6.13) \quad \begin{cases} (1,2),\ (1,3),\ \ldots,\ (1,n), \\ \qquad (2,3),\ \ldots,\ (2,n), \\ \qquad\qquad \vdots \\ \qquad\qquad (n-1,n), \end{cases}$$

sofern sie von Null verschieden sind, gewählt; diese Reihenfolge wird dann zyklisch wiederholt. Die Konvergenz dieses Verfahrens – für die angegebene, nicht für eine beliebige Reihenfolge der Pivotelemente – wurde von Forsythe und Henrici [19] bewiesen.

Unter zusätzlichen Voraussetzungen sind das zyklische Jacobi-Verfahren und das Verfahren mit maximaler Pivotwahl schließlich quadratisch konvergent; hierzu verweisen wir auf Wilkinson [62] und Schönhage [48].

Bei einer Variante des zyklischen Verfahrens, der Threshold-Methode, gibt man sich eine streng monoton fallende Nullfolge $(\tau_N)_0^\infty$ in $\mathbb{R}$ von „Schranken" (thresholds) vor, beispielsweise

$$(5.6.14)\quad \begin{cases} \tau_0 = \frac{1}{n}\,|\tilde{A}|\,, \\ \tau_{N+1} = \frac{1}{n}\,\tau_N \qquad (N \in \mathbb{N})\,. \end{cases}$$

Man wählt nun für die ersten Jacobi-Schritte – beispielsweise in der Reihenfolge (5.6.13) – nur solche Pivotindizes, für die

$$|a^{(k)}_{p_k, q_k}| \geq \tau_0$$

gilt. Sobald dann kein $|a^{(k)}_{i,j}| \geq \tau_0$ $(i<j)$ mehr existiert, ersetzt man τ_0 durch τ_1, läßt also nur Pivotindizes mit $|a^{(k)}_{p_k, q_k}| \geq \tau_1$ zu, usf. Die Konvergenz dieses Verfahrens ist leicht einzusehen (Übungsaufgabe 5.8).

Wie bereits gesagt, bricht man das Jacobi-Verfahren ab, sobald mit einem geeigneten $\epsilon > 0$ $|\tilde{A}_k| < \epsilon$ erreicht ist. Nach (5.6.11) gilt dann mit der unitären Matrix $U_k := S_0 \dots S_k$

$$(5.6.15)\quad A_k = U_k^* A\, U_k\,;$$

insbesondere ist mithin A_k zu A ähnlich. Wegen

$$(5.6.16)\quad A_k = \operatorname{diag}(a^{(k)}_{1,1}, \dots, a^{(k)}_{n,n}) + \tilde{A}_k\,, \quad |\tilde{A}_k| < \epsilon$$

wählt man als Näherungen der Eigenwerte von A_k (also von A) die Eigenwerte von $\operatorname{diag}(a^{(k)}_{1,1}, \dots, a^{(k)}_{n,n})$, d.h. die $a^{(k)}_{i,i}$ $(i = 1, \dots, n)$. Die Matrix U_k in (5.6.15) genügt der Beziehung

$$(5.6.17)\quad AU_k = U_k A_k \approx U_k \cdot \operatorname{diag}(a^{(k)}_{1,1}, \dots, a^{(k)}_{n,n})\,,$$

daher betrachtet man ihre Spalten als Näherungen für Eigenvektoren von A. Fehlerabschätzungen hierzu werden wir im folgenden Abschnitt bringen.

Man kann übrigens zeigen, daß die A_k (nicht nur bis auf Vertauschungen in der Diagonalen) gegen eine Diagonalmatrix konvergieren, ein Beweis dazu ist von Wilkinson [61], S. 268 angegeben.

Das Jacobi-Verfahren hat den Vorteil, daß es sich einfach programmieren läßt; insbesondere gewinnt man durch Berechnung von U_k in (5.6.15) sehr bequem ein Orthonormalsystem aus angenäherten Eigenvektoren zu A.

Andererseits konvergiert das Jacobi-Verfahren, da die quadratische Konvergenz erst spät eintritt, recht langsam. Die Zahl der Rechenoperationen läßt sich nicht dadurch verringern, daß man A in eine Tridiagonalmatrix transformiert, da die Tridiagonalgestalt durch Jacobi-Schritte zerstört wird.

Daher sollte man das Jacobi-Verfahren nur bei kleinem n und nur dann, wenn auch Eigenvektoren zu A gesucht sind, anwenden; sonst ist das in Abschnitt 5.5 behandelte QR-Verfahren wegen seiner i.a. erheblich kürzeren Rechenzeiten vorzuziehen.

(5.6.18) **Zahlenbeispiel.** Wir wenden das Threshold-Jacobi-Verfahren mit den τ_N aus (5.6.14) auf die in Beispiel (5.4.16) angegebene Matrix A an. Nach 12 Iterationen gewinnen wir A_{12} mit

$$\tilde{A}_{12} = 0{,}269 \cdot 10^{-9}$$

und, auf mehr als 10 Dezimalen hinter dem Komma genau,

$$\operatorname{diag}(a_{1,1}^{(12)}, \ldots, a_{4,4}^{(12)}) = \operatorname{diag}(15{,}00;\ -1{,}00;\ 5{,}00;\ 5{,}00)\,.$$

Die Matrix U_{12}, auf 10 Dezimalstellen gerundet, lautet

$$\begin{pmatrix} 0{,}4999999999 & -0{,}5000000000 & 0{,}0318505035 & -0{,}7063890893 \\ 0{,}5000000001 & 0{,}4999999998 & -0{,}7063890893 & -0{,}0318505035 \\ 0{,}5000000001 & 0{,}4999999999 & 0{,}7063890893 & 0{,}0318505035 \\ 0{,}4999999998 & -0{,}5000000002 & -0{,}0318505035 & 0{,}7063890893 \end{pmatrix}$$

Die Spalten von U_{12} stimmen bis auf 9 Dezimalen mit den exakten Eigenvektoren überein.

5.7. Fehlerbetrachtungen bei Eigenwertaufgaben

Es seien $A, S \in M(n \times n, \mathbb{C})$. Zu behandeln ist das folgende Problem: wie unterscheiden sich die Eigenwerte der Matrix A von denen der gestörten Matrix $A + S$?

Wir leiten zunächst eine qualitative Aussage dieser Art her. Hierzu bezeichnen wir mit $\lambda_1, \ldots, \lambda_n$ die Eigenwerte der Matrix A, mit $\lambda_1', \ldots, \lambda_n'$ die Eigenwerte der Matrix $A + S$, jeweils entsprechend ihrer Ordnung gezählt. $\|\ \|$ sei eine beliebige Matrixnorm in $M(n \times n, \mathbb{C})$.

(5.7.1) **Satz.** *Zu jedem $\eta > 0$ existiert ein $\delta > 0$ mit folgender Eigenschaft: ist $S \in M(n \times n, \mathbb{C})$ mit $\|S\| < \delta$, so gibt es eine Numerierung der Eigenwerte von $A + S$, so daß für jedes $\nu = 1, \ldots, n$*

$$|\lambda_\nu' - \lambda_\nu| < \eta$$

gilt.

Zum *Beweis* setzen wir

$$d = \begin{cases} \infty\,, & \text{falls A nur einen Eigenwert besitzt,} \\ \min\{|\lambda - \lambda'| : \lambda, \lambda' \text{ Eigenwerte von } A, \lambda \neq \lambda'\} & \text{sonst} \end{cases}$$

und nehmen im Fall $d < \infty$ o.E. $0 < \eta < \frac{d}{2}$ an. Offensichtlich ist die Menge

$$K = \{\mu \in \mathbb{C}: \exists \text{ Eigenwert } \lambda \text{ von } A \text{ mit } |\mu - \lambda| = \eta\}$$

kompakt und daher mit $\varphi(\mu) := \det(\mu I - A)$ auf Grund der Definition von d

$$\tau := \min\{|\varphi(\mu)| : \mu \in K\} > 0 .$$

Nach (3.3.3) existiert ein $\gamma \geq 0$, so daß für jedes $S = (s_{i,j})_{(n,n)} \in M(n \times n, \mathbb{C})$

$$\max_{i,j=1}^{n} |s_{i,j}| \leq \gamma |S|$$

abschätzbar ist. Da die Koeffizienten von $\varphi(\mu)$ stetig von A abhängen, existiert infolgedessen ein $\delta > 0$, so daß für jedes S mit $|S| < \delta$ das Polynom

$$\psi(\mu) := \det(\mu I - (A + S))$$

die Eigenschaft

$$\max_{\mu \in K} |\varphi(\mu) - \psi(\mu)| < \tau$$

besitzt.

Es sei nun λ ein (beliebiger) Eigenwert von A; dieser habe die Ordnung q, d.h. λ sei eine q-fache Nullstelle des Poylnoms $\varphi(\mu)$. In

$$K_\eta(\lambda) := \{\mu \in \mathbb{C}: |\mu - \lambda| < \eta\}$$

liegt außer λ kein weiterer Eigenwert von A. Wegen

$$\partial K_\eta(\lambda) = \{\mu \in \mathbb{C}: |\mu - \lambda| = \eta\} \subseteq K$$

gilt

$$|\varphi(\mu) - \psi(\mu)| < |\varphi(\mu)| \qquad (\mu \in \partial K_\eta(\lambda)) .$$

Hieraus folgt mit dem Satz von Rouché, daß in $K_\eta(\lambda)$, entsprechend der Ordnung gezählt, genau q Nullstellen von $\psi(\mu)$, also q Eigenwerte von A + S liegen. Damit ist die Behauptung des Satzes (5.7.1) bewiesen, da die Summe der Ordnungen $q(\lambda)$ gerade n ergibt.

Eine erste quantitative Aussage, d.h. eine erste Eigenwertabschätzung enthält der folgende von Ostrowski [44] stammende, ohne Beweis zitierte

(5.7.2) **Satz.** *Es seien* $A = (a_{i,j})_{(n,n)}$, $S = (s_{i,j})_{(n,n)} \in M(n \times n, \mathbb{C})$, *ferner*

$$M := \max\{|a_{i,j}|, |a_{i,j} + s_{i,j}| : i,j = 1, \ldots, n\} > 0$$

sowie

$$\delta := \frac{1}{M} \sum_{i,j=1}^{n} |s_{i,j}| .$$

Dann existiert zu jedem Eigenwert λ' *von* $A+S$ *ein Eigenwert* λ *von* A *mit*

$$|\lambda' - \lambda| \leqslant (n+2)M \cdot \delta^{1/n} .$$

Außerdem lassen sich die Eigenwerte λ_ν *und* λ'_ν *von* A *bzw.* $A+S$ *entsprechend ihrer Ordnung so numerieren, daß*

$$|\lambda'_\nu - \lambda_\nu| \leqq 2(n+1)^2 M \cdot \delta^{1/n} \qquad (\nu = 1, \ldots, n)$$

erfüllt ist.

Es gibt Beispiele – vgl. Übungsaufgabe 5.9 –, in denen sich für $|S| \to 0$ die $|\lambda'_\nu - \lambda_\nu|$ tatsächlich wie $|S|^{1/n}$ verhalten. In den folgenden Überlegungen wollen wir durch zusätzliche Voraussetzungen an A ein Verhalten $|\lambda'_\nu - \lambda_\nu| = O(|S|)$ $(|S| \to 0)$ und damit schärfere Abschätzungen als in Satz (5.7.2) erreichen. Als Hilfsmittel verwenden wir den von Bauer/Fike [3] bewiesenen

(5.7.3) **Satz.** *Es seien* $B, C \in M(n \times n, \mathbb{C})$, $\lambda \in \mathbb{C}$ *Eigenwert von* C. *Dann ist entweder* λ *Eigenwert von* B, *oder es gilt für jede Matrixnorm in* $M(n \times n, \mathbb{C})$

$$|(\lambda I - B)^{-1}(C - B)| \geqq 1 . \tag{5.7.4}$$

Beweis. Wir nehmen an, es sei λ kein Eigenwert von B, also $(\lambda I - B)$ invertierbar. Für einen Eigenvektor $x \neq 0$ zum Eigenwert λ von C gilt offenbar die Gleichung

$$(C - B)x = (\lambda I - B)x$$

und folglich

$$(\lambda I - B)^{-1}(C - B)x = x .$$

Hieraus folgt (5.7.4) zunächst für jede Operatornorm – vgl. (3.3.4) – und wegen (3.3.7), (ii) auch für jede Matrixnorm in $M(n \times n, \mathbb{C})$.

Bevor wir das Störungsproblem weiter behandeln, wollen wir aus (5.7.3) zwei *Einschließungssätze* für die Eigenwerte einer Matrix folgern.

(5.7.5) **Satz** (Gerschgorin). *Ist* $A = (a_{i,j})_{(n,n)} \in M(n \times n, \mathbb{C})$, λ *Eigenwert von* A, *so gilt für mindestens ein* $i \in \{1, \ldots, n\}$ *die Ungleichung*

$$|\lambda - a_{i,i}| \leqq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}| . \tag{5.7.6}$$

Zum *Beweis* wenden wir Satz (5.7.3) auf $C = A$, $B = \operatorname{diag}(a_{1,1}, \ldots, a_{n,n})$ bezüglich der Matrixnorm $\| \|_\infty$ – vgl. (3.3.8) – an.

Es sei λ Eigenwert von A. Ist λ gleichzeitig Eigenwert von B, so existiert ein $i \in \{1, \ldots, n\}$ mit $\lambda = a_{i,i}$, und (5.7.6) gilt trivialerweise. Andernfalls liefert (5.7.4) die Ungleichung

$$1 \leq \max_{i=1}^{n} \left(\frac{1}{|\lambda - a_{i,i}|} \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}| \right)$$

und damit (5.7.6) für mindestens einen Index i.

Wir bezeichnen als *Gerschgorinkreise* die abgeschlossenen Kreisscheiben

$$(5.7.7) \qquad K_i := \left\{ \mu \in \mathbb{C}: |\mu - a_{i,i}| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}| \right\} \qquad (i = 1, \ldots, n).$$

Hiermit formulieren wir als *2. Satz von Gerschgorin* den

(5.7.8) **Satz.** *Es sei* $J := \{i_1, \ldots, i_r\}$ *eine Teilmenge von* $\{1, \ldots, n\}$; *es habe*

$$G := \bigcup_{i \in J} K_i$$

die Eigenschaft

$$G \cap K_j = \emptyset \qquad (j \in \{1, \ldots, n\} \setminus J).$$

Dann liegen in G, *entsprechend der Ordnung gezählt, genau* r *Eigenwerte von* A.

Zum *Beweis* nehmen wir o.E. $\emptyset \subsetneq J \subsetneq \{1, \ldots, n\}$ an. Dann haben wir, da die K_i kompakt sind,

$$(5.7.9) \qquad \operatorname{dist}\left(G, \bigcup_{j \notin J} K_j\right) =: \eta > 0.$$

Es sei nun mit $B := \operatorname{diag}(a_{1,1}, \ldots, a_{n,n})$

$$A(t) := B + t \cdot (A - B) \qquad (0 \leq t \leq 1)$$

definiert. Dann ist $A(0) = B$, $A(1) = A$; ferner gilt nach (5.7.6) für jeden Eigenwert $\lambda(t)$ von $A(t)$ $(t \in [0, 1])$ mindestens eine der Ungleichungen

$$|\lambda(t) - a_{i,i}| \leq t \cdot \sum_{j \neq i} |a_{i,j}| \leq \sum_{j \neq i} |a_{i,j}|,$$

folglich liegen sämtliche Eigenwerte von $A(t)$ in $\bigcup_{i=1}^{n} K_i$. Wenn wir die Eigenwerte hier wie im Folgenden entsprechend ihrer Ordnung zählen, so enthält G trivialerweise genau r Eigenwerte von $A(0) = B$. Sinnvoll ist daher die Definition

(5.7.10) $t_0 := \sup \{t \in [0, 1]$: genau r Eigenwerte von $A(t)$ liegen in $G\}$;

$\lambda_1(t_0), \ldots, \lambda_n(t_0)$ seien die Eigenwerte von $A(t_0)$. Nach Satz (5.7.1) existiert ein $\delta > 0$ und weiter zu jedem (festen) $t \in [0, 1]$ mit $|t - t_0| < \delta$ eine Numerierung der Eigenwerte $\lambda_1(t), \ldots, \lambda_n(t)$ von $A(t)$, so daß

$$(5.7.11) \qquad |\lambda_i(t) - \lambda_i(t_0)| < \eta \qquad (i = 1, \ldots, n) .$$

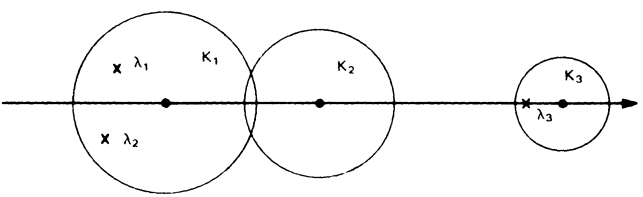
Gemäß der Definition von t_0 liegen für mindestens ein $t \in [0, 1]$ mit $t_0 - \delta < t \leqq t_0$ genau r der $\lambda_i(t)$ in G; folglich enthält nach (5.7.9) und (5.7.11) G ebenfalls genau r der $\lambda_i(t_0)$, mithin ist t_0 das Maximum der in (5.7.10) angegebenen Menge.

Es bleibt zu zeigen, daß $t_0 = 1$ ist; dann ist wegen $A(1) = A$ der Beweis abgeschlossen. Unter der Annahme $t_0 < 1$ gäbe es ein t mit $t_0 < t \leqq 1$, für das (5.7.11) erfüllt wäre. Für ein solches t lägen wiederum genau r der $\lambda_i(t)$ in G, im Widerspruch zur Definition (5.7.10).

Ist K_{i_0} ein Gerschgorinkreis mit der Eigenschaft

$$K_{i_0} \cap K_i = \emptyset \qquad (i \neq i_0) ,$$

so kann man den Satz (5.7.8) auf $G = K_{i_0}$ anwenden; folglich liegt in K_{i_0} genau ein Eigenwert von A. Wenn andernfalls G eine zusammenhängende Menge aus $r \geqq 2$ Gerschgorin-Kreisen ist, so lassen sich die in G liegenden Eigenwerte von A nicht den $K_i \subset G$ eineindeutig zuordnen; vielmehr ist eine Verteilung wie in folgender Skizze möglich:



Wir kommen zur Anwendung von Satz (5.7.3) auf das anfangs erwähnte Störungsproblem. Hierbei betrachten wir als Matrixnormen nur Operatornormen, die die zusätzliche Eigenschaft

$$(5.7.12) \qquad \forall D = \operatorname{diag}(d_1, \ldots, d_n) \quad |D| = \max_{i=1}^{n} |d_i|$$

besitzen. Offenbar ist (5.7.12) für die Operatornormen zu sämtlichen in (3.2.3) angegebenen Normen des $\mathbb{C}^n$ erfüllt. Man kann zeigen (vgl. Bauer, Stoer, Witzgall [5]), daß (5.7.12) der Bedingung

$$\forall x \in \mathbb{C}^n \qquad |x| = |\hat{x}|$$

– zur Definition von $\hat{x}$ siehe (3.5.5) – äquivalent ist.

Unter der Voraussetzung (5.7.12) gilt der

(5.7.13) **Satz.** *Es seien* $A, S \in M(n \times n, \mathbb{C})$; *zu* A *gebe es eine invertierbare Matrix* T *und eine Diagonalmatrix* D *mit*

$$A = TDT^{-1}. \tag{5.7.14}$$

Dann existiert zu jedem Eigenwert λ' *von* $A+S$ *ein Eigenwert* λ *von* A *mit*

$$|\lambda' - \lambda| \leqq |T|\,|T^{-1}|\,|S| = \kappa(T)\,|S|. \tag{5.7.15}$$

Beweis. Wir wenden Satz (5.7.3) auf $B := A$, $C := A+S$ an. Ist ein Eigenwert λ' von $A+S$ gleichzeitig Eigenwert von A, so ist (5.7.15) trivialerweise erfüllt. Andernfalls gilt nach (5.7.4)

$$1 \leqq |(\lambda' I - A)^{-1} S| \leqq |(\lambda' I - A)^{-1}|\,|S|.$$

Gemäß (5.7.14) haben wir die Beziehung

$$(\lambda' I - A)^{-1} = (T(\lambda' I - D)T^{-1})^{-1} = T(\lambda' I - D)^{-1} T^{-1},$$

woraus wir die Abschätzung

$$1 \leqq |(\lambda' I - D)^{-1}|\,|T|\,|T^{-1}|\,|S|,$$

also

$$\frac{1}{|(\lambda' I - D)^{-1}|} \leqq |T|\,|T^{-1}|\,|S| \tag{5.7.16}$$

folgern. Da D die Eigenwerte von A als Diagonalelemente besitzt, ergibt sich unter Berücksichtigung von (5.7.12)

$$|(\lambda' I - D)^{-1}| = \max\left\{ \frac{1}{|\lambda' - \lambda|} : \lambda \text{ Eigenwert von } A \right\},$$

mithin

$$\frac{1}{|(\lambda' I - D)^{-1}|} = \min\{|\lambda' - \lambda| : \lambda \text{ Eigenwert von } A\}.$$

Dies in (5.7.16) eingesetzt, liefert die Behauptung.

Wir gewinnen als

(5.7.17) **Folgerung.** Ist $A \in M(n \times n, \mathbb{C})$ normal, d.h. $A^*A = AA^*$, $S \in M(n \times n, \mathbb{C})$ beliebig, so existiert zu jedem Eigenwert λ' von $A+S$ ein Eigenwert λ von A mit

$$|\lambda' - \lambda| \leqq \|S\|_S \leqq \|S\|_2. \tag{5.7.18}$$

Hierbei bezeichne $\|\ \|_S$ die Operatornorm bezüglich der euklidischen Norm, $\|\ \|_2$ die Frobeniusnorm (vgl. (3.3.8)).

Zum *Beweis* beachten wir, daß A die Bedingung (5.7.14) mit einer unitären Matrix T erfüllt; es ist also bezüglich der euklidischen Norm $\kappa(T) = 1$.

Unter den Voraussetzungen von Satz (5.7.13) seien $\lambda_1, \ldots, \lambda_n$ die Eigenwerte von A, entsprechend ihrer Ordnung gezählt. Mit den Bezeichnungen

(5.7.19) $\quad K(\lambda_i) := \{\mu \in \mathbb{C}: |\mu - \lambda_i| \leqq \kappa(T)\,|S|\} \qquad (i = 1, \ldots, n)$

notieren wir den

(5.7.20) **Zusatz.** Ist $J := \{i_1, \ldots, i_r\}$ eine Teilmenge von $\{1, \ldots, n\}$ und hat

$$G := \bigcup_{i \in J} K(\lambda_i)$$

die Eigenschaft

$$G \cap K(\lambda_j) = \emptyset \qquad (j \in \{1, \ldots, n\} \setminus J),$$

so liegen in G, entsprechend der Ordnung gezählt, genau r Eigenwerte von $A + S$.

Zum *Beweis* setzen wir

$$A(t) := A + tS \qquad (0 \leqq t \leqq 1)$$

und verfahren wie im Beweis zu (5.7.8).

Es sei G in (5.7.20) zusammenhängend; mit $\lambda'_{i_1}, \ldots, \lambda'_{i_r}$ seien die in G enthaltenen Eigenwerte von $A + S$ bezeichnet. Dann gelten, wie man sich leicht überlegt, die Abschätzungen

(5.7.21) $\quad |\lambda'_i - \lambda_i| \leqq (2r - 1)\,\kappa(T)\,|S| \leqq (2n - 1)\,\kappa(T)\,|S| \qquad (i \in J).$

Unser nächstes Ziel ist eine schärfere Abschätzung als (5.7.21) für den Fall, daß A und S hermitesch sind. Als Hilfsmittel benutzen wir den

(5.7.22) **Satz** *(Minimum-Maximum-Prinzip). Es sei* $A \in M(n \times n, \mathbb{C})$ *hermitesch. Wir numerieren die (reellen) Eigenwerte von* A *entsprechend ihrer Ordnung so, daß*

$$\lambda_1 \leqq \lambda_2 \leqq \ldots \leqq \lambda_n$$

gilt. Dann ist für $\nu = 1, \ldots, n$

$$\lambda_\nu = \min \left\{ \max_{x \in U,\, \neq 0} \frac{(Ax, x)}{(x, x)} : U \textit{ Unterraum des } \mathbb{C}^n,\ \dim U = \nu \right\}.$$

Beweis. Es sei $(v_1, \ldots, v_n)$ eine Orthonormalbasis des $\mathbb{C}^n$ aus Eigenvektoren v_i zu den Eigenwerten λ_i. Hierzu bezeichnen wir

$$E_0 := \{0\}, \quad E_\nu := \operatorname{span}(v_1, \ldots, v_\nu) \qquad (\nu = 1, \ldots, n).$$

Zunächst behaupten wir:

(α) Für jeden Unterraum U des $\mathbb{C}^n$ mit $\dim U = \nu$ gilt

$$\max_{x \in U,\, \neq 0} \frac{(Ax, x)}{(x, x)} \geqq \lambda_\nu.$$

Zum Beweis von (α) notieren wir für

$$E_{\nu-1}^{\perp} = \operatorname{span}(v_\nu, \ldots, v_n)$$

auf Grund des Dimensionssatzes die Eigenschaft

$$\begin{aligned}\dim(U \cap E_{\nu-1}^{\perp}) &= \dim U + \dim E_{\nu-1}^{\perp} - \dim(U + E_{\nu-1}^{\perp}) \\ &\geqq \nu + n - \nu + 1 - n = 1 ,\end{aligned}$$

folglich existiert ein $x \in U \cap E_{\nu-1}^{\perp}$, $\neq 0$. Ein solches x besitzt die Darstellung

$$x = \sum_{i=\nu}^{n} \alpha_i v_i \qquad (\alpha_i \in \mathbb{C}) ;$$

daher gilt

$$Ax = \sum_{i=\nu}^{n} \lambda_i \alpha_i v_i$$

und weiter, da die v_i orthonormiert sind,

$$(x, x) = \sum_{i=\nu}^{n} |\alpha_i|^2 ,$$

$$(Ax, x) = \sum_{i=\nu}^{n} \lambda_i |\alpha_i|^2 \geqq \lambda_\nu \sum_{i=\nu}^{n} |\alpha_i|^2 = \lambda_\nu (x, x) .$$

Hiermit ist (α) nachgewiesen. Als zweite Aussage, die mit (α) zusammen die Behauptung des Satzes ergibt, zeigen wir

(β) Es existiert ein Unterraum U des $\mathbb{C}^n$ mit $\dim U = \nu$ und

$$\max_{x \in U, \neq 0} \frac{(Ax, x)}{(x, x)} \leqq \lambda_\nu .$$

Hierzu wählen wir $U := E_\nu$. Dann gilt für beliebiges $x \in U$, $\neq 0$ eine Darstellung

$$x = \sum_{i=1}^{\nu} \alpha_i v_i .$$

Es folgt

$$(Ax, x) = \sum_{i=1}^{\nu} \lambda_i |\alpha_i|^2 \leqq \lambda_\nu \sum_{i=1}^{\nu} |\alpha_i|^2 = \lambda_\nu (x, x) ,$$

mithin, wie behauptet,

$$\max_{x \in U, \neq 0} \frac{(Ax, x)}{(x, x)} \leqq \lambda_\nu .$$

Hieraus folgern wir den

(5.7.23) **Satz.** *Es seien* $A, S \in M(n \times n, \mathbb{C})$ *hermitesch;* $\lambda_1 \leqq \lambda_2 \leqq \ldots \leqq \lambda_n$ *seien die Eigenwerte von* A, $\sigma_1 \leqq \ldots \leqq \sigma_n$ *die Eigenwerte von* S *und schließlich* $\lambda'_1 \leqq \ldots \leqq \lambda'_n$ *die Eigenwerte von* A + S, *jeweils entsprechend ihrer Ordnung gezählt. Dann gelten die Abschätzungen*

(i) $\lambda_\nu + \sigma_1 \leqq \lambda'_\nu \leqq \lambda_\nu + \sigma_n \qquad (\nu = 1, \ldots, n)$,

und folglich für jede Matrixnorm

(ii) $|\lambda_\nu - \lambda'_\nu| \leqq |S| \qquad (\nu = 1, \ldots, n)$.

Beweis. Ist U ein beliebiger Unterraum des $\mathbb{C}^n$ mit $\dim U = \nu$, so folgt aus dem vorangehenden Satz

$$\lambda'_\nu \leqq \max_{x \in U, \neq 0} \frac{((A+S)x, x)}{(x, x)} \leqq \max_{x \in U, \neq 0} \frac{(Ax, x)}{(x, x)} + \max_{x \in U, \neq 0} \frac{(Sx, x)}{(x, x)} .$$

Wir wählen speziell $U = E_\nu = \operatorname{span}(v_1, \ldots, v_\nu)$, wobei die v_i orthonormierte Eigenvektoren zu den λ_i von A sind. Dann gilt – vgl. den Beweisteil (β) von (5.7.22) –

$$\max_{x \in U, \neq 0} \frac{(Ax, x)}{(x, x)} = \lambda_\nu$$

und selbstverständlich

$$\max_{x \in U, \neq 0} \frac{(Sx, x)}{(x, x)} \leqq \max_{x \in \mathbb{C}^n, \neq 0} \frac{(Sx, x)}{(x, x)} = \sigma_n .$$

Hiermit haben wir

(5.7.24) $\lambda'_\nu \leqq \lambda_\nu + \sigma_n \qquad (\nu = 1, \ldots, n)$

gezeigt. Wenn wir die entsprechenden Überlegungen auf A + S statt A, $-S$ statt S anwenden, so gewinnen wir, da $-S$ die Eigenwerte

$$-\sigma_n \leqq -\sigma_{n-1} \leqq \ldots \leqq -\sigma_1$$

besitzt, aus (5.7.24) für die Eigenwerte von $(A + S) - S = A$ die Ungleichungen

$$\lambda_\nu \leqq \lambda'_\nu - \sigma_1 \qquad (\nu = 1, \ldots, n),$$

insgesamt also die Aussage (i) des Satzes. Aussage (ii) folgt wegen

$$|\sigma_1|, \ |\sigma_n| \leqq \rho(S) \leqq |S|$$

– vgl. (3.3.11) – unmittelbar aus (i).

Es ergibt sich für das Jacobi-Verfahren die

(5.7.25) **Folgerung.** Es sei $A \in M(n \times n, \mathbb{R})$ symmetrisch, A_k wie in (5.6.16) zu A unitär ähnlich. Dann gelten bei einer geeigneten Anordnung der Eigenwerte λ_i

von A entsprechend ihrer Ordnung die Abschätzungen

$$|\lambda_i - a_{i,i}^{(k)}| \leq |\tilde{A}_k| \quad (< \epsilon) \qquad (i = 1, \ldots, n) .$$

Die bisherigen Überlegungen lassen sich ebenfalls für eine Fehlerrechnung beim QR-Verfahren verwenden: dies wird in der Übungsaufgabe 5.11 diskutiert. Ferner läßt sich ein mit der Potenzmethode berechneter Eigenwert einer normalen Matrix A gemäß Übungsaufgabe 5.12 abschätzen.

Abschließend wollen wir Fehlerbetrachtungen für Eigenvektoren durchführen. Hierzu nehmen wir an, $A \in M(n \times n, \mathbb{C})$ sei normal (beispielsweise hermitesch); $\lambda_1, \ldots, \lambda_n$ seien die Eigenwerte von A, entsprechend ihrer Ordnung gezählt, $(v_1, \ldots, v_n)$ sei eine zugehörige Orthonormalbasis des $\mathbb{C}^n$ aus Eigenvektoren von A.

Dann bezeichnen wir für $J = \{i_1, \ldots, i_r\} \subseteq \{1, \ldots, n\}$

$$E(J) := \operatorname{span}(v_{i_1}, \ldots, v_{i_r}) .$$

Ist beispielsweise λ ein Eigenwert von A und

$$J = \{j \in \{1, \ldots, n\} : \lambda_j = \lambda\} , \tag{5.7.26}$$

so ist selbstverständlich $E(J) = E(\lambda)$, also der Eigenraum zu λ. Wir verwenden im Folgenden die euklidische Norm im $\mathbb{C}^n$ und die zugehörige Operatornorm (Spektralnorm) in $M(n \times n, \mathbb{C})$. Für $x \in \mathbb{C}^n$, U Unterraum des $\mathbb{C}^n$ sei der Abstand von x zum Unterraum U wie üblich als

$$\operatorname{dist}(x, U) := \inf \{|x - y| : y \in U\}$$

erklärt. Unter diesen Voraussetzungen gilt der

(5.7.27) **Satz.** *Es sei* $x \in \mathbb{C}^n, \neq 0$, $\tilde{\lambda} \in \mathbb{C}$, J *eine nichtleere Teilmenge von* $\{1, \ldots, n\}$ *mit der Eigenschaft*

$$|\tilde{\lambda} - \lambda_j| \geq d > 0 \qquad (j \in \{1, \ldots, n\} \setminus J) .$$

Wir behaupten

$$\operatorname{dist}(x, E(J)) \leq \frac{1}{d} |Ax - \tilde{\lambda} x| .$$

Beweis. Es sei o.E. $J = \{1, \ldots, r\}$ $(1 \leq r \leq n)$. Wenn wir

$$(*) \qquad x = \sum_{i=1}^{n} \alpha_i v_i \qquad (\alpha_i \in \mathbb{C})$$

darstellen, so ergibt sich, da die v_i orthonormierte Eigenvektoren zu A sind, die Beziehung

$$(**) \qquad |Ax - \tilde{\lambda} x|^2 = \left| \sum_{i=1}^{n} (\lambda_i - \tilde{\lambda}) \alpha_i v_i \right|^2 = \sum_{i=1}^{n} |\lambda_i - \tilde{\lambda}|^2 |\alpha_i|^2 .$$

Außerdem gewinnen wir mit (*) eine Zerlegung

$$x = x_I + x_{II}$$

in

$$x_I := \sum_{i=1}^{r} \alpha_i v_i \in E(J), \quad x_{II} := \sum_{i=r+1}^{n} \alpha_i v_i \in E(J)^{\perp},$$

die nach Definition des Abstandes zu der Ungleichung

$$(***) \qquad \operatorname{dist}(x, E(J)) \leq |x - x_I| = |x_{II}|$$

führt (wobei wegen $x_{II} \in E(J)^{\perp}$ sogar Gleichheit eintritt). Weiter erhalten wir auf Grund der Voraussetzung

$$|\tilde{\lambda} - \lambda_j| \geq d \qquad (j = r+1, \ldots, n)$$

die Abschätzung

$$|x_{II}|^2 = \sum_{i=r+1}^{n} |\alpha_i|^2 \leq \frac{1}{d^2} \sum_{i=r+1}^{n} |\lambda_i - \tilde{\lambda}|^2 \, |\alpha_i|^2 \leq \frac{1}{d^2} \sum_{i=1}^{n} |\lambda_i - \tilde{\lambda}|^2 \, |\alpha_i|^2$$

und mit (***) und (**) schließlich die Behauptung.

Ist $\tilde{\lambda}$ Näherung zu einem Eigenwert λ von A, x näherungsweise ein zugehöriger Eigenvektor und J durch (5.7.26) gegeben, so liefert der vorstehende Satz eine Abschätzung für den Abstand von x zum Eigenraum $E(\lambda)$. Diese Abschätzung wird unbrauchbar, wenn d sehr klein wird, also in der Nähe von $\tilde{\lambda}$ (bzw. λ) weitere Eigenwerte $\lambda_i \neq \lambda$ von A liegen. In solchen Fällen wählt man

$$J = \{i \in \{1, \ldots, n\} : \lambda_i \approx \tilde{\lambda}\},$$

betrachtet also den Abstand von x zur direkten Summe der Eigenräume, die zu den in der Nähe von $\tilde{\lambda}$ liegenden Eigenwerten gehören.

Wir gewinnen aus Satz (5.7.27) die

(5.7.28) **Folgerung.** Es sei $S \in M(n \times n, \mathbb{C})$, ferner $x \in \mathbb{C}^n$ mit $|x| = 1$ Eigenvektor zum Eigenwert λ' von $A + S$; schließlich J eine nichtleere Teilmenge von $\{1, \ldots, n\}$ mit der Eigenschaft

$$|\lambda' - \lambda_j| \geq d > 0 \qquad (j \in \{1, \ldots, n\} \setminus J).$$

Dann gilt

$$\operatorname{dist}(x, E(J)) \leq \frac{1}{d} |S|.$$

Zum *Beweis* beachten wir

$$|Ax - \lambda' x| = |Sx| \leq |S|.$$

Als Anwendung leiten wir eine Fehlerabschätzung für die Spalten der im Jacobi-Verfahren gewonnenen unitären Matrizen U_k mit $A_k = U_k^* A U_k$ – vgl. (5.6.15) – her. Hierzu schreiben wir

$$D_k := \operatorname{diag}(a_{1,1}^{(k)}, \ldots, a_{n,n}^{(k)})$$

und benutzen (5.6.15), woraus sich

$$A = U_k A_k U_k^* = U_k D_k U_k^* + U_k \tilde{A}_k U_k^* ,$$

also

$$U_k D_k U_k^* = A + T_k$$

mit

$$T_k := - U_k \tilde{A}_k U_k^*$$

ergibt. Dabei haben wir – vgl. Hilfssatz (5.6.2) –

$$|T_k| = |\tilde{A}_k| ,$$

ferner besitzt $A + T_k$ zu den Eigenwerten $a_{i,i}^{(k)}$ die Eigenvektoren $U_k e_i$ $(i = 1, \ldots, n)$. Ist daher $i \in \{1, \ldots, n\}$ fest und J eine nichtleere Teilmenge von $\{1, \ldots, n\}$ mit der Eigenschaft

$$|\lambda_j - a_{i,i}^{(k)}| \geqq d > 0 \qquad (j \in \{1, \ldots, n\} \setminus J) ,$$

so folgt aus (5.7.28) unmittelbar

$$\text{(5.7.29)} \qquad \operatorname{dist}(U_k e_i , E(J)) \leqq \frac{1}{d} |\tilde{A}_k| .$$

Weitergehende störungstheoretische Untersuchungen, insbesondere über analytische Störungen, finden sich u.a. in der umfassenden Monographie von Kato [36].

Übungsaufgaben zum 5. Kapitel

Aufgabe 5.1. Man zeige – unter der Voraussetzung von Satz (5.1.10) – die abgeschwächte Aussage, daß in der Darstellung

$$y_k = \lambda_1^k (h_1 + v_k) \qquad (k \in \mathbb{N})$$

die v_k mit einem $\gamma \geqq 0$ und gewissen $\epsilon_k \geqq 0$ mit $\lim\limits_{k \to \infty} \epsilon_k = 0$ den Abschätzungen

$$\|v_k\|_\infty \leqq \gamma \cdot [\rho(1 + \epsilon_k)]^k$$

genügen. Hierzu benutze man an Stelle des Hilfssatzes (5.1.9) die Aussage (6.2.17) gemäß folgender

Anleitung: Man transformiert

$$S^{-1}AS = J = \begin{pmatrix} J_1 & 0 \\ 0 & J_2 \end{pmatrix} ,$$

wobei J_1 den Eigenwert λ_1, J_2 die Eigenwerte $\lambda_2, \ldots, \lambda_m$ besitzt. Man setzt

$$y_0 =: Sz_0, \; z_0 = \begin{pmatrix} z_0^1 \\ z_0^2 \end{pmatrix} ,$$

letzteres gemäß der Aufteilung von J. Dann ist $h_1 = S \begin{pmatrix} z_0^1 \\ 0 \end{pmatrix}$, folglich $z_0^1 \neq 0$ Eigenvektor zum Eigenwert λ_1 von J_1. Man betrachtet nun die Folge $(z_k)_0^\infty$ mit

$$z_k = J^k z_0 \quad (k \in \mathbb{N})$$

und wendet (6.2.17) auf J_2 an.

Aufgabe 5.2. Man beweise (5.2.17), (ii), indem man aus der Annahme $h_1 = 0$ die Tatsache $h_2 = h_3 = \ldots = h_m = 0$ und damit einen Widerspruch herleitet. Dazu definiert man

$$n' := n - q_1 \; (\leqq n-1); \; P_{n'}(\lambda) := (\lambda - \lambda_1)^{-q_1} P(\lambda)$$

und betrachtet die durch Streichen der letzten q_1 Komponenten aus den h_i entstehenden Vektoren $h_i' \in \mathbb{C}^{n'}$ $(i = 2, \ldots, n)$. Dann sind die h_i' Hauptvektoren der nach (5.2.7) zu $P_{n'}$ gehörenden Matrix, und es gilt

$$h_i = 0 \iff h_i' = 0 \quad (i = 2, \ldots, n) .$$

Aufgabe 5.3. Es sei $P(\lambda)$ wie in (5.2.8) gegeben; wir bezeichnen

$$B(\lambda) := \lambda \cdot P'(\lambda) - n \cdot P(\lambda) .$$

Für $k = 0, 1, 2, \ldots$ sei

$$G_k(\lambda) = \alpha_{n-1,k}\lambda^{n-1} + \alpha_{n-2,k}\lambda^{n-2} + \ldots + \alpha_{0,k}$$

der Rest bei der Division von $\lambda^k B(\lambda)$ durch $P(\lambda)$, also durch

$$\lambda^k \cdot B(\lambda) = P(\lambda) \cdot Q_k(\lambda) + G_k(\lambda)$$

definiert. Schließlich sei

$$\eta_k := \alpha_{n-1,k} \quad (k \in \mathbb{N}) .$$

(i) Man zeige

$$G_0(\lambda) = B(\lambda), \; G_{k+1}(\lambda) = \lambda \cdot G_k(\lambda) - \eta_k P(\lambda) \quad (k \in \mathbb{N}) .$$

(ii) Man bestimme die $\alpha_{i,0}$ und leite Rekursionen für die $\alpha_{i,k}$ her.

(iii) Man folgere aus (ii) für die η_k die Beziehungen

$$\begin{cases} \eta_0 = -a_{n-1}, \\ \eta_k = -\left[(k+1)a_{n-k-1} + \sum_{i=0}^{k-1} a_{n-k+i}\eta_i\right] & (k = 1, \ldots, n-1), \\ \eta_k = -\sum_{i=1}^{n} a_{n-i}\eta_{k-i} & (k = n, n+1, \ldots), \end{cases}$$

d.h. die η_k erfüllen die Differenzengleichung (5.2.6) mit den Anfangswerten (5.2.17), (iii).

(iv) Man zeige

$$G_k(\lambda) = \sum_{i=1}^{m} q_i \frac{P(\lambda)}{(\lambda - \lambda_i)} \lambda_i^{k+1}; \quad \eta_k = \sum_{i=1}^{m} q_i \lambda_i^{k+1} \qquad (k \in \mathbb{N}).$$

(v) Unter der Voraussetzung $|\lambda_1| > |\lambda_2| \geqq \ldots \geqq |\lambda_m|$ gilt mit $\rho := \dfrac{|\lambda_2|}{|\lambda_1|}$

a) $$\frac{\eta_{k+1}}{\eta_k} = \lambda_1 + O(\rho^k) \quad (k \to \infty);$$

b) $$\frac{1}{\eta_k} G_k(\lambda) = \frac{1}{\lambda - \lambda_1} P_k(\lambda) + O(\rho^k) \quad (k \to \infty)$$

gleichmäßig in kompakten Teilmengen von $\mathbb{C}$.

Anleitung: Zu (iii) zeigt man mit Induktion über k

$$\alpha_{i,k} = -\left[(k+n-i)a_{i-k} + \sum_{j=0}^{k-1} a_{i+1+j-k}\eta_j\right] \qquad (i = 0, \ldots, n-1),$$

wobei $a_\nu = 0$ für $\nu < 0$ zu setzen ist; zu (iv) leite man im Fall $k = 0$ aus

$$P(\lambda) = \prod_{i=1}^{m} (\lambda - \lambda_i)^{q_i}$$

eine geeignete Darstellung von $P'(\lambda)$ her.

Aufgabe 5.4. Im Beweis des Satzes (5.3.4) bezeichnen wir

$$P_1 := I, \; P_\nu := P_{\nu, i_\nu} \ldots P_{2, i_2} \qquad (\nu = 2, \ldots, n-1)$$

mit $P_{\mu, i_\mu} := I$, falls beim μ-ten Konstruktionsschritt der Fall (I) eintritt. Wir definieren $\pi_\nu \in S_n$, $\pi := \pi_{n-1}$ durch

$$e^t_{\pi_\nu(i)} = e_i^t P_\nu \qquad (i = 1, \ldots, n)$$

und setzen

$$A^{(\nu)} =: (\alpha^{(\nu)}_{\pi_\nu(i), \pi_\nu(j)})_{(n,n)} \qquad (\nu = 1, \ldots, n-1).$$

(i) Man formuliere einen Algorithmus – analog (2.2.16) – zur Bestimmung von $\pi_{\nu+1}$ und Berechnung der $\alpha^{(\nu+1)}_{\pi_{\nu+1}(i),\,\pi_{\nu+1}(j)}$ $(\nu = 1, \ldots, n-2)$. Hierbei verwende man die Größen

$$d_{\pi_{\nu+1}(i),\,\pi_{\nu+1}(\nu)} = \begin{cases} 0 \text{ im Fall (I), d.h. } \alpha^{(\nu)}_{\pi_{\nu+1}(\nu+1),\,\pi_{\nu+1}(\nu)} = 0\,, \\ \dfrac{\alpha^{(\nu)}_{\pi_{\nu+1}(i),\,\pi_{\nu+1}(\nu)}}{\alpha^{(\nu)}_{\pi_{\nu+1}(\nu+1),\,\pi_{\nu+1}(\nu)}} \qquad \text{sonst} \\ (i = \nu+2, \ldots, n)\,. \end{cases}$$

(ii) Man zeige: die Matrix L in (5.3.9) besitzt die Koeffizienten

$$l_{i,j} = \begin{cases} 1 & (i = j)\,, \\ d_{\pi(i),\,\pi(j-1)} & (j+1 \leqq i \leqq n;\ 2 \leqq j \leqq n)\,, \\ 0 \quad \text{sonst}\,. \end{cases}$$

(iii) Es sei $T = P^{-1}L$ wie in (5.3.9), $x = (\xi_i)_1^n \in \mathbb{C}^n$ beliebig und hierzu

$$y = (\eta_i)_1^n := Tx\,.$$

Man gebe eine formelmäßige Darstellung der η_i unter Benutzung der $d_{\pi(i),\pi(j)}$.

Aufgabe 5.5. Es sei $A \in M(n \times n, \mathbb{C})$ hermitesch; $A^{(1)} = A$ sei unterteilt wie im Beweis von Satz (5.3.11), hierbei gelte $f_1 \neq 0$. Wir setzen – mit den Bezeichnungen (5.3.12) –

$$u := f_1 - \sigma_1 \mu_1 e_1\,,\quad N := \mu_1^2 + \mu_1 |a_{2,1}|\,,\quad v := \frac{1}{N}u\,,$$

$$p^* := u^* E_1\,,\quad s := p - \tfrac{1}{2}(p^* u)v\,.$$

Man zeige:

(i) $G_1 E_1 G_1 = E_1 - vs^* - sv^*$.

(ii) Die Berechnung von $H_1 A H_1$ erfordert – nach (i) unter Berücksichtigung der Symmetrie – 1 Quadratwurzel sowie $[2(n-1)^2 + 5n - 4]$ Multiplikationen bzw. Divisionen.

(iii) Die Berechnung der Tridiagonalmatrix $C = Q^* A Q$ gemäß Satz (5.3.11) erfordert $(n-1)$ Wurzeln und etwa $\frac{2}{3}n^3$ Multiplikationen.

Aufgabe 5.6. Es sei $A_0 \in M(n \times n, \mathbb{C})$ invertierbare obere Hessenbergmatrix, ferner Q_0 unitär, R_0 obere Dreiecksmatrix mit

$$A_0 = Q_0 R_0\,.$$

Hiermit sei

$$A_1 := R_0 Q_0$$

gesetzt.

Man zeige:

(i) Q_0 und A_1 haben obere Hessenbergform;

(ii) ist A_0 hermitesche Tridiagonalmatrix, so ist auch A_1 hermitesche Tridiagonalmatrix.

Aufgabe 5.7. Es sei $A \in M(n \times n, \mathbb{C})$; hierzu seien A_k $(k \in \mathbb{N})$ gemäß der Vorschrift (5.5.35) mit gewissen $\mu_k \in \mathbb{C}$ konstruiert.

Man zeige für $k \in \mathbb{N}$, ≥ 1:

(i) $A_k = (Q_0 \dots Q_{k-1})^* A\,(Q_0 \dots Q_{k-1})$,

(ii) $(Q_0 \dots Q_{k-1})(R_{k-1} \dots R_0) = (A - \mu_0 I)(A - \mu_1 I) \dots (A - \mu_{k-1} I)$,

(iii) $A_k = (R_{k-1} \dots R_0)\, A\, (R_{k-1} \dots R_0)^{-1}$;

(iv) falls A die Voraussetzungen (5.5.3), (i), (ii) erfüllt, so erfüllt auch A_k die Voraussetzungen (5.5.3), (i), (ii).

Aufgabe 5.8. Man zeige die Konvergenz des Threshold-Jacobi-Verfahrens.

Aufgabe 5.9. Zu

$$A := \begin{pmatrix} 0 & 1 & 0 & \dots & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & 0 & \\ \vdots & & & \ddots & 1 & \\ 0 & \dots & \dots & \dots & \dots & 0 \end{pmatrix}_{(n,n)}$$

bestimme man $S \in M(n \times n, \mathbb{C})$, so daß für alle Eigenwerte $\lambda'_\nu(t)$ von $A + tS$ die Gleichungen

$$|\lambda'_\nu(t) - \lambda'_\nu(0)| = |\lambda'_\nu(t)| = |t|^{\frac{1}{n}} \qquad (\nu = 1, \dots, n;\ t \in \mathbb{C})$$

gelten.

Aufgabe 5.10. Es sei

$$P(z) = a_n z^n + a_{n-1} z^{n-1} + \dots + a_0, \quad a_n = 1$$

komplexes Polynom; $\lambda \in \mathbb{C}$ sei eine Nullstelle von P. Man zeige die Ungleichungen

(i) $$|\lambda| \leq \max \left\{ 1, \sum_{i=0}^{n-1} |a_i| \right\},$$

(ii) $$|\lambda| \leq \max \{|a_0|,\ 1 + |a_1|, \dots, 1 + |a_{n-1}|\}$$

und im Fall $a_1 \cdot \ldots \cdot a_{n-1} \neq 0$

(iii) $|\lambda| \leqq \max \left\{ \frac{|a_0|}{|a_1|}, \; 2\frac{|a_1|}{|a_2|}, \ldots, 2\frac{|a_{n-1}|}{|a_n|} \right\}$,

(iv) $|\lambda| \leqq \sum_{i=0}^{n-1} \frac{|a_i|}{|a_{i+1}|}$.

Anleitung: Man wendet Satz (5.7.5) auf eine geeignete Matrix A an. In den Fällen (iii) und (iv) konstruiert man A durch Ähnlichkeitstransformation mit einer Diagonalmatrix.

Aufgabe 5.11. Es sei $A \in M(n \times n, \mathbb{C})$, A erfülle die Voraussetzung (5.5.3), (i). Als Norm im $\mathbb{C}^m$ $(m \leqq n)$ sei die euklidische Norm gewählt.

(i) Auf A wird das QR-Verfahren (5.5.35) mit gewissen Parametern $\mu_k \in \mathbb{C}$ angewendet. Nach k_1 Schritten habe man

$$A_{k_1} = \left(\begin{array}{c|c} A'_{k_1} & b_{n-1}^{(k_1)} \\ \hline a_{n-1}^{(k_1)t} & a_{n,n}^{(k_1)} \end{array} \right)$$

mit

$$A'_{k_1} \in M(n-1 \times n-1, \mathbb{C}), \; b_{n-1}^{(k_1)}, \; a_{n-1}^{(k_1)} \in \mathbb{C}^{n-1} .$$

Man zeige: es existiert ein Eigenwert λ von A mit

$$|\lambda - a_{n,n}^{(k_1)}| \leqq |a_{n-1}^{(k_1)}| \cdot \kappa(T) ,$$

wobei $\kappa(T)$ bezüglich der euklidischen Operatornorm gemeint ist.

Hinweis: Wegen Aufgabe 5.7, (i) erfüllt A_{k_1} die Voraussetzung (5.5.3), (i) mit einer Matrix T', für die $\kappa(T) = \kappa(T')$ gilt.

(ii) Nach $(k_2 - k_1)$ QR-Schritten (5.5.35), auf A'_{k_1} angewendet, erhalte man

$$A'_{k_2} = \left(\begin{array}{c|c} A''_{k_2} & b_{n-2}^{(k_2)} \\ \hline a_{n-2}^{(k_2)t} & a_{n-1,n-1}^{(k_2)} \end{array} \right)$$

mit

$$A''_{k_2} \in M(n-2 \times n-2, \mathbb{C}), \; a_{n-2}^{(k_2)}, \; b_{n-2}^{(k_2)} \in \mathbb{C}^{n-2} .$$

Man gebe eine unitäre Matrix $\tilde{Q}$ an, so daß gilt

$$\tilde{Q}^* A_{k_1} \tilde{Q} = \left(\begin{array}{c|c} A'_{k_2} & b_{n-1}^{(k_2)} \\ \hline a_{n-1}^{(k_2)t} & a_{n,n}^{(k_1)} \end{array} \right) =: A_{k_2}$$

und hierbei

$$|a_{n-1}^{(k_2)}| = |a_{n-1}^{(k_1)}|, \quad |b_{n-1}^{(k_2)}| = |b_{n-1}^{(k_1)}| .$$

Man folgere: es existiert eine unitäre (n,n)-Matrix $Q^{(k_2)}$ mit

$$A_{k_2} = Q^{(k_2)*} A\, Q^{(k_2)} ;$$

ferner gibt es einen Eigenwert λ' von A mit

$$|\lambda' - a_{n-1,n-1}^{(k_2)}| \leqq (|a_{n-1}^{(k_1)}|^2 + |a_{n-2}^{(k_2)}|^2)^{\frac{1}{2}} .$$

(iii) Auf A_{k_2}'' werden wiederum $(k_3 - k_2)$ QR-Schritte (5.5.35) angewendet usf. Durch analoge Überlegungen wie zu (ii) gewinnt man hierzu schließlich

$$A_{k_{n-1}} = Q^{(k_{n-1})*} A\, Q^{(k_{n-1})} ,$$

wobei $Q^{(k_{n-1})}$ unitär ist und $A_{k_{n-1}}$ folgende Eigenschaften besitzt:

$$A_{k_{n-1}} = R - S$$

mit

$$R = (r_{i,j})_{(n,n)} = \begin{pmatrix} a_{1,1}^{(k_{n-1})} & & & & \\ 0 & a_{2,2}^{(k_{n-1})} & & & \\ \vdots & & \ddots & & \\ & & & a_{n-1,n-1}^{(k_2)} & \\ 0 & \cdots & \cdots & 0 & a_{n,n}^{(k_1)} \end{pmatrix} ,$$

$S = (s_{i,j})_{(n,n)}$ untere Dreiecksmatrix mit $s_{i,i} = 0$ $(i = 1, \ldots, n)$ und

$$|S|^2 \leqq \|S\|_2^2 = \sum_{\kappa=1}^{n-1} |a_{n-\kappa}^{(k_\kappa)}|^2 .$$

Man folgere aus (5.7.15): für alle $i = 1, \ldots, n$ existiert ein Eigenwert $\lambda^{(i)}$ von A mit

$$|r_{i,i} - \lambda^{(i)}| = \kappa(T)\,|S| \leqq \kappa(T)\,\|S\|_2 .$$

(iv) Ist A und damit $A_{k_{n-1}}$ hermitesch, so gilt bezüglich einer geeigneten Anordnung $\lambda_{\nu_1}, \ldots, \lambda_{\nu_n}$ der Eigenwerte von A

$$|r_{i,i} - \lambda_{\nu_i}| \leqq \left(2 \sum_{\kappa=1}^{n-1} |a_{n-\kappa}^{(k_\kappa)}|^2\right)^{\frac{1}{2}} \qquad (i = 1, \ldots, n) .$$

Hinweis: Man setzt

$$A_{k_{n-1}} = \operatorname{diag}(r_{1,1}, \ldots, r_{n,n}) + \tilde{S} .$$

Aufgabe 5.12. Es sei $A \in M(n \times n, \mathbb{C})$ normal, $\tilde{\lambda} \in \mathbb{C}$, ferner $x \in \mathbb{C}^n$ mit $|x| = 1$, wobei die euklidische Norm im $\mathbb{C}^n$ gemeint sei. Man zeige

$$\min \{|\tilde{\lambda} - \lambda| : \lambda \text{ Eigenwert von } A\} \leqq |Ax - \tilde{\lambda}x| .$$

Anleitung: Man gehe wie in Satz (5.7.27) vor.

Aufgabe 5.13. Für $t \in [-1, 1]$ sei

$$A(t) = \begin{pmatrix} t & -t & 0 \\ -1 & t & 0 \\ 0 & 0 & 1-t \end{pmatrix} .$$

(i) Man berechne die Eigenwerte von $A(t)$ und numeriere sie so, daß sie bezüglich t stetig sind.

(ii) Man untersuche das Verhalten der Eigenwerte sowie der Haupt- und Eigenräume von $A(t)$ in der Umgebung von $t = 0$ und $t = \frac{1}{4}$.

6. Iterationsverfahren

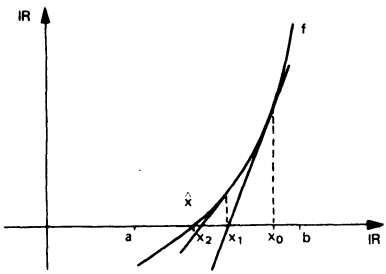
Vorgegeben sei über $[a, b]$ eine reellwertige Funktion f. In $\hat{x} \in]a, b[$ habe sie eine Nullstelle; eine Näherung hierzu sei x_0. Ist f zumindest in x_0 differenzierbar, so ist durch

$$t(x) = f(x_0) + f'(x_0)\,(x - x_0) \qquad (x \in \mathbb{R})$$

die Tangente an f im Punkte $(x_0, f(x_0))$ gegeben. Der Schnittpunkt dieser Tangente mit der x-Achse, d.h. die Bedingung $t(x_1) = 0$ liefert eventuell eine bessere Näherung x_1 von $\hat{x}$. Wiederholte Anwendung dieser Prozedur führt zum Newton-Verfahren (in $\mathbb{R}$)

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \qquad (n = 0, 1, 2, \ldots)\,.$$

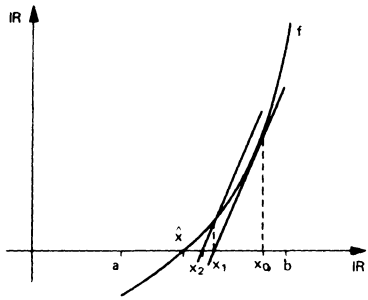
Unter geeigneten, in diesem Kapitel präzisierten Voraussetzungen konvergiert die so definierte Folge $(x_n)_0^\infty$ gegen die Nullstelle $\hat{x}$.



Newton-Verfahren

Da die numerische Berechnung der Ableitungen $f'(x_n)$ häufig Schwierigkeiten bereitet, betrachtet man statt der Tangenten an f in $(x_n, f(x_n))$ die Geraden durch $(x_n, f(x_n))$ jeweils mit der gleichen Steigung $f'(x_0)$. Auf diese Weise erhält man das vereinfachte Newton-Verfahren (in $\mathbb{R}$)

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)} \qquad (n = 0, 1, 2, \ldots)\,.$$



Vereinfachtes Newton-Verfahren

Ist $f'(x) \neq 0$ in $[a,b]$, so gilt mit

$$T(x) := x - \frac{f(x)}{f'(x)} \qquad (x \in [a,b])$$

die Beziehung

$$T(\hat{x}) = \hat{x} \iff f(\hat{x}) = 0 ;$$

ferner erfüllt die durch das Newton-Verfahren definierte Folge $(x_n)_0^\infty$ die Iterationsvorschrift

$$x_{n+1} = T(x_n) \qquad (n = 0, 1, 2, \ldots) .$$

Im Fall des vereinfachten Newton-Verfahrens setzt man entsprechend

$$T(x) := x - \frac{f(x)}{f'(x_0)} \qquad (x \in [a,b]) .$$

Die Aufgabe, eine Nullstelle von f zu bestimmen, ist damit auf ein Fixpunktproblem zurückgeführt. Derartige Fixpunktprobleme werden wir im nächsten Abschnitt eingehend behandeln.

6.1. Der Banachsche Fixpunktsatz

Es sei in diesem Abschnitt stets (R, d) ein metrischer Raum. Weiter sei T eine Abbildung von D_T (Definitionsbereich von T) $\subset R$ in R; dabei lassen wir auch $D_T = \emptyset$ zu.

Rekursiv definieren wir die „Potenzen" T^n des Operators T.

(6.1.1) **Definition.** Es sei

$$\begin{cases} D_{T^0} := R, \quad T^0 = id_R ; \\ D_{T^{n+1}} := \{x \in D_T : T(x) \in D_{T^n}\}, \quad T^{n+1}(x) := T^n(T(x)) \quad (x \in D_{T^{n+1}}) \\ (n = 0, 1, 2, \ldots) . \end{cases}$$

Hierzu notieren wir den

(6.1.2) **Hilfssatz.** *Für* $x \in R$ *hat man*

$$x \in D_{T^{n+m}} \iff x \in D_{T^m} \quad \textit{und} \quad T^m(x) \in D_{T^n} ;$$

entsprechend gilt für $x \in D_{T^{n+m}}$

$$T^{n+m}(x) = T^n(T^m(x)) .$$

Beweis. Wir schließen induktiv: Für $m = 0$ ist die Aussage trivialerweise gültig. Weiter sei $x \in D_{T^{n+m+1}}$. Gemäß (6.1.1) tritt dies genau dann ein, wenn $x \in D_T$, $T(x) \in D_{T^{n+m}}$ sowie $T^{n+m+1}(x) = T^{n+m}(T(x))$ gilt. Nach Induktionsannahme bedeutet $T(x) \in D_{T^{n+m}}$ gerade $T(x) \in D_{T^m}$, $T^m(T(x)) \in D_{T^n}$ und

$$T^{n+m}(T(x)) = T^n(T^m(T(x))) .$$

Beachtet man noch, daß abermals nach (6.1.1) $x \in D_T$, $T(x) \in D_{T^m}$ genau dann, wenn $x \in D_{T^{m+1}}$ und $T^{m+1}(x) = T^m(T(x))$, so ergibt sich die Gültigkeit der Aussage des Hilfssatzes für $m + 1$.

D_T ist bezüglich der Einschränkung der Metrik d selbst metrischer Raum; demgemäß bezeichnet man diese Einschränkung als (die von d auf D_T erzeugte) *Relativmetrik.* Im Folgenden sei T im Sinne der Definition (3.1.16) als Abbildung von D_T (versehen mit der Relativmetrik) in R beschränkt. Gemäß (3.1.16) ist dann $|T|$ erklärt.

(6.1.3) **Hilfssatz.** *Für* $n \in \mathbb{N}$ *sind die Potenzen* T^n *beschränkt; weiter gelten für* $n, m \in \mathbb{N}$ *die Abschätzungen*

$$|T^{n+m}| \leqq |T^n| \, |T^m| , \tag{6.1.4}$$

$$|T^{nm}| \leqq |T^n|^m . \tag{6.1.5}$$

Darüber hinaus hat man

$$\sum_{n=0}^{\infty} |T^n| < \infty \iff \exists\, p \in \mathbb{N}, \neq 0 \quad |T^p| < 1 .$$

Beweis. Durch Induktion beweist man die Beschränktheit der Operatoren T^n; dabei zeigt man unter Benutzung von (3.1.20) die Abschätzung

$$|T^n| \leqq |T|^n \quad (n \in \mathbb{N}) . \tag{6.1.6}$$

Ebenfalls induktiv und unter Anwendung von (3.1.20) und (6.1.2) weist man die Gültigkeit der Abschätzungen (6.1.4), (6.1.5) nach.

Ist die Reihe $\sum_{n=0}^{\infty} |T^n|$ konvergent, so strebt $|T^n| \to 0$ für $n \to \infty$. Daher existiert in diesem Fall sicher ein $p \in \mathbb{N}, \neq 0$ mit $|T^p| < 1$.

Umgekehrt habe man nun ein $p \in \mathbb{N}, \neq 0$ mit $|T^p| < 1$. Für beliebiges $k \in \mathbb{N}$ schätzt man mit (6.1.4) und (6.1.5)

$$\sum_{n=0}^{kp-1} |T^n| = \sum_{l=0}^{k-1} \left(\sum_{i=0}^{p-1} |T^{pl+i}| \right)$$

$$\leqq \sum_{l=0}^{k-1} |T^p|^l \sum_{i=0}^{p-1} |T^i| \leqq \frac{1}{1-|T^p|} \sum_{i=0}^{p-1} |T^i|$$

ab, womit die Konvergenz der Reihe $\sum_{n=0}^{\infty} |T^n|$ aufgezeigt ist.

Nach diesen Vorbereitungen beweisen wir den für viele Bereiche der Angewandten Mathematik grundlegenden

(6.1.7) **Satz** *(Fixpunktsatz von Banach-Weissinger). Es sei* (R,d) *ein vollständiger metrischer Raum,* $x_0 \in R$, $0 < \rho \leqq \infty$, $0 \leqq \sigma < \infty$. *Es bezeichne*

$$K(x_0, \rho) := \{x \in R : d(x, x_0) \,(\leqq)\, \rho\}.$$

T *sei eine beschränkte Abbildung von* $D_T = K(x_0, \rho)$ *in* R *mit*

$$\sum_{n=0}^{\infty} |T^n| \leqq \sigma \tag{6.1.8}$$

und

$$\sigma \cdot d(x_0, T(x_0)) \,(\leqq)\, \rho. \tag{6.1.9}$$

Wir behaupten:

(i) T *besitzt genau einen Fixpunkt* $\hat{x}$ *in* $K(x_0, \rho)$, *d.h. es existiert genau ein* $\hat{x} \in K(x_0, \rho)$ *mit* $T(\hat{x}) = \hat{x}$.

(ii) *Für alle* $n \in \mathbb{N}$ *ist* $x_0 \in D_{T^n}$; *setzt man* $x_n := T^n(x_0)$, *so genügt die Folge* $(x_n)_0^{\infty}$ *der Iterationsvorschrift* $x_{n+1} = T(x_n)$ *und konvergiert gegen* $\hat{x}$ *für* $n \to \infty$.

(iii) *Es gelten die a priori Abschätzung*

$$d(x_n, \hat{x}) \leqq \left(\sum_{\nu=n}^{\infty} |T^\nu| \right) \cdot d(x_1, x_0) \quad (n \in \mathbb{N}) \tag{6.1.10}$$

und die a posteriori Abschätzung

$$d(x_n, \hat{x}) \leqq \left(\sum_{\nu=1}^{\infty} |T^\nu| \right) d(x_n, x_{n-1}) \quad (n \in \mathbb{N}, \neq 0). \tag{6.1.11}$$

Beweis. a) Wir zeigen durch Induktion nach n:

$$\forall n \in \mathbb{N} \quad x_0 \in D_{T^n},\ d(x_0, x_n) \leqq \sum_{\nu=0}^{n-1} |T^\nu| \cdot d(x_0, x_1) .$$

Für $n = 0$ ist diese Aussage klar. Wir kommen zum Schluß von n auf $n+1$. Wegen (6.1.8) und (6.1.9) hat man nach Induktionsvoraussetzung $x_0 \in D_{T^n}$, $T^n(x_0) \in D_T = K(x_0, \rho)$, folglich gemäß Hilfssatz (6.1.2) $x_0 \in D_{T^{n+1}}$ und

$$x_{n+1} = T^{n+1}(x_0) = T(x_n) = T^n(x_1) .$$

Unter Benutzung der Abschätzung in der Induktionsvoraussetzung ergibt sich dann noch

$$\begin{aligned} d(x_0, x_{n+1}) &\leqq d(x_0, x_n) + d(x_n, x_{n+1}) \\ &\leqq \sum_{\nu=0}^{n-1} |T^\nu| \cdot d(x_0, x_1) + |T^n| \cdot d(x_0, x_1) \\ &= \sum_{\nu=0}^{n} |T^\nu| \cdot d(x_0, x_1) . \end{aligned}$$

b) Als nächstes beweisen wir die Aussage

$$\exists \hat{x} \in K(x_0, \rho) \quad x_n \to \hat{x} \quad (n \to \infty) .$$

Für $k, n \in \mathbb{N}$ schätzt man

$$\begin{aligned} d(x_n, x_{n+k}) &\leqq \sum_{\nu=n}^{n+k-1} d(x_\nu, x_{\nu+1}) = \sum_{\nu=n}^{n+k-1} d(T^\nu(x_0), T^\nu(x_1)) \\ &\leqq \sum_{\nu=n}^{n+k-1} |T^\nu| \cdot d(x_0, x_1) , \end{aligned}$$

mithin

$$(6.1.12) \qquad d(x_n, x_{n+k}) \leqq \sum_{\nu=n}^{\infty} |T^\nu| \cdot d(x_0, x_1)$$

ab. Da $\sum_{\nu=n}^{\infty} |T^\nu| \to 0$ strebt für $n \to \infty$, ist die Folge $(x_n)_0^\infty$ C-konvergent in (R, d). Nach Voraussetzung ist dieser Raum vollständig; daher existiert ein $\hat{x} \in R$ mit $x_n \to \hat{x}$ $(n \to \infty)$. Aus (6.1.12) folgt

$$d(x_n, \hat{x}) \leqq \sum_{\nu=n}^{\infty} |T^\nu| \cdot d(x_0, x_1) + d(x_{n+k}, \hat{x}) ,$$

also mit $k \to \infty$ für $\hat{x}$ die Ungleichung (6.1.10). Für $n = 0$ liefert diese unter Beachtung von (6.1.9)

$$d(x_0, \hat{x}) \leqq \sigma \cdot d(x_0, T(x_0)) \overset{(\leqq)}{<} \rho ,$$

wonach, wie zu zeigen war, $\hat{x}$ sogar in $K(x_0, \rho)$ liegt.

c) Nun zur Aussage: $\hat{x}$ ist Fixpunkt, und zwar einziger Fixpunkt von T. Da T eine beschränkte, mithin stetige Abbildung von $K(x_0, \rho)$ in R ist, folgt aus $x_{n+1} = T(x_n)$ durch Grenzübergang $n \to \infty$ unmittelbar $\hat{x} = T(\hat{x})$. Dies kann man übrigens auch der Ungleichung

$$\begin{aligned} d(\hat{x}, T(\hat{x})) &\leqq d(\hat{x}, x_{n+1}) + d(T(x_n), T(\hat{x})) \\ &\leqq d(\hat{x}, x_{n+1}) + |T| \cdot d(x_n, \hat{x}), \to 0 \quad (n \to \infty) \end{aligned}$$

entnehmen.

Ist $\tilde{x} \in K(x_0, \rho)$ ebenfalls Fixpunkt von T, d.h. $T(\tilde{x}) = \tilde{x}$, so ergibt sich für $n \in \mathbb{N}$

$$T^n(\hat{x}) = \hat{x}, \; T^n(\tilde{x}) = \tilde{x} ,$$

mithin

$$d(\hat{x}, \tilde{x}) \leqq |T^n| \, d(\hat{x}, \tilde{x}) , \to 0 \quad (n \to \infty) ,$$

also, wie behauptet $\hat{x} = \tilde{x}$.

d) Die a priori Abschätzung ist bereits bewiesen. Zur Herleitung der a posteriori Abschätzung schließen wir in ähnlicher Weise: Für $k, n \in \mathbb{N}$ mit $n \geqq 1$ hat man

$$\begin{aligned} d(x_n, x_{n+k}) &\leqq \sum_{\nu=1}^{k} d(x_{n+\nu-1}, x_{n+\nu}) = \sum_{\nu=1}^{k} d(T^\nu(x_{n-1}), T^\nu(x_n)) \\ &\leqq \sum_{\nu=1}^{k} |T^\nu| \cdot d(x_{n-1}, x_n) , \end{aligned}$$

folglich

$$d(x_n, x_{n+k}) \leqq \sum_{\nu=1}^{\infty} |T^\nu| \, d(x_n, x_{n-1}) ,$$

mithin nach dem Grenzübergang $k \to \infty$ die behauptete Abschätzung.

Besonders hervorzuheben sind die folgenden Spezialfälle:

1. $\rho = \infty$, d.h. $K(x_0, \rho) = R$. In diesem Fall ist T eine beschränkte Abbildung von R in sich mit (6.1.8); die „technische" Voraussetzung (6.1.9) entfällt hier natürlich. Die Behauptung (ii) ist so zu verstehen, daß für jedes $x_0 \in R$ die Iterierten $x_n = T^n(x_0)$ für $n \to \infty$ gegen den Fixpunkt $\hat{x} \in R$ streben.

2. T kontrahierend, d.h. $|T| < 1$. Unter dieser Voraussetzung ist

$$\sum_{n=0}^{\infty} |T^n| \leqq \sum_{n=0}^{\infty} |T|^n = \frac{1}{1-|T|}$$

abschätzbar und daher $\sigma := \frac{1}{1-|T|}$ wählbar.

3. $\rho = \infty$, T kontrahierend. Hier liegt die ursprünglich von Banach [2] bewiesene Fassung des Fixpunktsatzes, des Prinzips der kontrahierenden Abbildung vor; sie lautet

(6.1.7′) **Satz.** *Es sei* (R, d) *ein vollständiger metrischer Raum,* T *eine kontrahierende Abbildung von* R *in sich.*

Behauptung:

(i) T *besitzt in* R *genau einen Fixpunkt* $\hat{x}$.

(ii) *Für jedes* $x_0 \in R$ *strebt* $x_n = T^n(x_0) \to \hat{x}$ *für* $n \to \infty$.

(iii) *Es gelten die a priori Abschätzung*

$$(6.1.10') \quad d(x_n, \hat{x}) \leqq \frac{|T|^n}{1-|T|}\, d(x_1, x_0) \quad (n \in \mathbb{N})$$

und die a posteriori Abschätzung

$$(6.1.11') \quad d(x_n, \hat{x}) \leqq \frac{|T|}{1-|T|}\, d(x_n, x_{n-1}) \quad (n \in \mathbb{N}, \neq 0).$$

(6.1.13) **Zahlenbeispiel.** Gesucht sind reelle Zahlen ξ_1, ξ_2, die das nichtlineare Gleichungssystem

$$\begin{cases} (1 + 0{,}05\,\xi_2)(1 - \xi_1) - 0{,}025\,\xi_2^2 = 0 \\ 0{,}025\,(1-\xi_1)^2 - \xi_2\,(1 - 0{,}02\,\xi_2) = 0{,}5 \end{cases}$$

erfüllen. Zur Anwendung des Fixpunktsatzes bringen wir dieses Gleichungssystem auf die Form

$$\begin{cases} \xi_1 = 1 + 0{,}05\,(1-\xi_1)\cdot\xi_2 - 0{,}025\,\xi_2^2 \\ \xi_2 = -0{,}5 + 0{,}025\,(1-\xi_1)^2 + 0{,}02\,\xi_2^2\,, \end{cases}$$

die eine Lösung in der Nähe des Punktes $(\xi_1; \xi_2) = (1; -0{,}5)$ vermuten läßt. Da für $\xi_2 < -0{,}5$ sicher keine Lösung existiert, ist es sinnvoll,

$$R := [0{,}5;\, 1{,}5] \times [-0{,}5;\, 0]$$

zu wählen.

Definiert man für $x = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \in R$

$$T(x) := \begin{pmatrix} 1 + 0{,}05\,(1-\xi_1)\,\xi_2 - 0{,}025\,\xi_2^2 \\ -0{,}5 + 0{,}025\,(1-\xi_1)^2 + 0{,}02\,\xi_2^2 \end{pmatrix} ,$$

so ist, wie man leicht nachprüft, T eine Abbildung von R in sich.

Für $x = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$, $y = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \in R$ sei – vgl. Definition (3.1.6) –

$$d(x,y) := d_\infty(x,y) = \max_{i=1;2} |\xi_i - \eta_i| ,$$

also d die von d_∞ auf R erzeugte Metrik. (R, d) ist vollständig, da $\mathbb{R}^2$ bezüglich d_∞ vollständig und R eine (bezüglich d_∞) abgeschlossene Teilmenge von $\mathbb{R}^2$ ist; letzteres bedeutet, daß $x_n \in R$, $\to x \in \mathbb{R}^2$ für $n \to \infty$ die Aussage $x \in R$ impliziert. Bezüglich d ist T eine kontrahierende Abbildung; genauer gilt

$$d(T(x), T(y)) \leqq 0{,}05\,(|\xi_1 - \eta_1| + |\xi_2 - \eta_2|) \leqq 0{,}1\; d(x,y) ,$$

das heißt

$$|T| \leqq 0{,}1 .$$

Zur Iteration stützen wir uns – es ist $\rho = \infty$, T kontrahierend – auf den Banachschen Fixpunktsatz (6.1.7′). Als Startwert bietet sich

$$x_0 = \begin{pmatrix} 1 \\ -0{,}5 \end{pmatrix}$$

an. Wir benutzen die im Beispiel (1.1.16) angegebene REAL*8-Arithmetik, die etwa einer 16-stelligen dezimalen Gleitkomma-Arithmetik entspricht. Bei exakter Rechnung erhielten wir beim ersten Iterationsschritt $\bar{x}_1 := T(x_0)$; statt dessen berechnen wir $\tilde{x}_1 := gl(T(x_0))$. Diesen Wert $\tilde{x}_1$ nehmen wir als neuen Startwert und berechnen hiermit $\tilde{x}_2 = gl(T(\tilde{x}_1))$ anstelle von $\bar{x}_2 := T(\tilde{x}_1)$ usw. Zur Abschätzung des Abbrechfehlers nach dem i-ten Iterationsschritt wenden wir (6.1.10′) bzw. (6.1.11′) an, und zwar für n = 1, indem wir $x_0 := \tilde{x}_{i-1}$ und $x_1 := \bar{x}_i$ setzen. Dies führt für i = 1, 2, 3, ... zu

$$d(\bar{x}_i, \hat{x}) \leqq \frac{0{,}1}{1-0{,}1}\, d(\bar{x}_i, \tilde{x}_{i-1}) \leqq 0{,}112\; d(\bar{x}_i, \tilde{x}_{i-1}) .$$

Mittels der Überlegungen aus Kapitel 1 erhält man ohne Mühe mit $\tau = \frac{1}{2}\, 10^{-15}$ ebenfalls für i = 1, 2, 3, ...

$$d(\tilde{x}_i, \bar{x}_i) \leqq 1{,}1\,\tau$$

und daher insgesamt

$$\begin{aligned} d(\tilde{x}_i, \hat{x}) &\leqq 1{,}1\,\tau + 0{,}112\,(1{,}1\,\tau + d(\tilde{x}_i, \tilde{x}_{i-1})) \\ &\leqq 0{,}612 \cdot 10^{-15} + 0{,}112\; d(\tilde{x}_i, \tilde{x}_{i-1}) . \end{aligned}$$

Im Folgenden sind für $i = 1, 2, 3, \ldots$ die Werte für $\tilde{x}_i$ und für die Abschätzungen von $d(\tilde{x}_i, \hat{x})$ angegeben.

i	$\tilde{x}_i$		$d(\tilde{x}_i, \hat{x}) \leqq$
0	1,0000000000000000	−0,5000000000000000	
1	0,9937500000000001	−0,4950000000000000	$0{,}700 \cdot 10^{-3}$
2	0,9937196875000002	−0,4950985234375000	$0{,}111 \cdot 10^{-4}$
3	0,9937164676299764	−0,4950965629836727	$0{,}361 \cdot 10^{-6}$
4	0,9937164370690542	−0,4950966007969590	$0{,}424 \cdot 10^{-8}$
5	0,9937164353645824	−0,4950966000385083	$0{,}191 \cdot 10^{-9}$
6	0,9937164353414019	−0,4950966000529930	$0{,}260 \cdot 10^{-11}$
7	0,9937164353404651	−0,4950966000526989	$0{,}106 \cdot 10^{-12}$
8	0,9937164353404493	−0,4950966000527044	$0{,}239 \cdot 10^{-14}$
9	0,9937164353404487	−0,4950966000527043	$0{,}700 \cdot 10^{-15}$
10	0,9937164353404487	−0,4950966000527043	$0{,}612 \cdot 10^{-15}$

6.2. Iterationsverfahren bei linearen Problemen

Wir notieren zunächst den

(6.2.1) **Satz.** *Es sei* $(R, |\ |)$ *ein Banach-Raum über* $\mathbb{K} = \mathbb{R}$ *oder* $\mathbb{C}$, *ferner* $b \in R$ *und* $A \in \mathrm{Hom}(R)$ *mit*

$$(6.2.2) \qquad \sum_{m=0}^{\infty} |A^m| < \infty .$$

Wir behaupten:

(i) *Es existiert genau ein* $\hat{x} \in R$ *mit*

$$(6.2.3) \qquad A\hat{x} + b = \hat{x} .$$

(ii) *Für alle* $x_0 \in R$ *strebt die durch die Vorschrift*

$$(6.2.4) \qquad x_{m+1} = Ax_m + b \quad (m \in \mathbb{N})$$

definierte Folge $(x_m)_0^{\infty}$ *für* $m \to \infty$ *gegen* $\hat{x}$.

(iii) *Es gelten die a priori Abschätzung*

$$(6.2.5) \qquad |x_m - \hat{x}| \leqq \left(\sum_{\nu=m}^{\infty} |A^{\nu}| \right) |x_1 - x_0| \quad (m \in \mathbb{N})$$

und die a posteriori Abschätzung

$$(6.2.6) \qquad |x_m - \hat{x}| \leqq \left(\sum_{\nu=1}^{\infty} |A^{\nu}| \right) |x_m - x_{m-1}| \quad (m \in \mathbb{N}, \neq 0) ,$$

die sich unter der zusätzlichen Voraussetzung $|A| < 1$ *zu*

$$(6.2.5') \qquad |x_m - \hat{x}| \leqq \frac{|A|^m}{1-|A|}\,|x_1 - x_0| \qquad (m \in \mathbb{N})$$

beziehungsweise

$$(6.2.6') \qquad |x_m - \hat{x}| \leqq \frac{|A|}{1-|A|}\,|x_m - x_{m-1}| \qquad (m \in \mathbb{N}, \neq 0)$$

vereinfachen.

Zum *Beweis* wenden wir den Fixpunktsatz (6.1.7) bezüglich $\rho = \infty$ an. Hierzu betrachten wir R mit der durch die Norm erzeugten Metrik und definieren durch

$$T(x) := Ax + b \qquad (x \in R)$$

eine Abbildung von R in sich. Für $m \in \mathbb{N}$ gilt

$$T^m(x) = A^m x + \sum_{\nu=0}^{m-1} A^\nu b \qquad (x \in R)\,,$$

wie man durch Induktion leicht nachweist. Hiermit erhält man für $x, y \in R$

$$T^m(x) - T^m(y) = A^m x - A^m y = A^m(x-y)\,,$$

also $|T^m| = |A^m|$ und daher, wie zu zeigen war,

$$\sum_{m=0}^{\infty} |T^m| < \infty\,.$$

Aussage (i) folgt übrigens auch bereits aus Satz (3.2.13); danach ist nämlich $(I-A)$ bijektiv und $(I-A)^{-1} \in L(R)$.

Die Beschränktheit von $(I-A)^{-1}$ wollen wir zur Herleitung zusätzlicher Abschätzungen des Abbrechfehlers benutzen: Gemäß (6.2.3) und (6.2.4) hat man einerseits

$$(I-A)(\hat{x} - x_m) = b - x_m + x_{m+1} - b = x_{m+1} - x_m$$

und andererseits

$$x_{m+1} - x_m = A(x_m - x_{m-1}) = \ldots = A^m(x_1 - x_0)\,.$$

Hiermit erhält man als a priori bzw. a posteriori Abschätzung

$$(6.2.7) \qquad |x_m - \hat{x}| \leqq |(I-A)^{-1}A^m|\,|x_1 - x_0| \leqq |(I-A)^{-1}|\,|A^m|\,|x_1 - x_0|\,,$$

$$(6.2.8) \qquad |x_m - \hat{x}| \leqq |(I-A)^{-1}A|\,|x_m - x_{m-1}| \leqq |(I-A)^{-1}|\,|A|\,|x_m - x_{m-1}|\,.$$

Ein Iterationsverfahren der Form (6.2.4) – mit $A \in \mathrm{Hom}(R)$ – nennen wir im Folgenden konvergent, falls die hierdurch definierte Folge $(x_m)_0^\infty$ bei jedem Startwert $x_0 \in R$ und jeder „Inhomogenität" $b \in R$ konvergiert. Der Grenzwert $\hat{x}$ einer derartigen konvergenten Folge erfüllt dann, falls $A \in L(R)$ ist, natürlich die Gleichung (6.2.3).

Den vorstehenden Satz spezialisieren wir weiter, und zwar u.a. insbesondere auf den Fall linearer Gleichungssysteme.

(6.2.9) **Satz.** *Es sei* $\mathbb{K} = \mathbb{R}$ *oder* $\mathbb{C}$, $A \in M(n \times n, \mathbb{K})$, $= (a_{i,j})_{(n,n)}$, *ferner* $b \in \mathbb{K}^n$ *sowie* $w = (w_i)_1^n \in \mathbb{R}^n$, > 0 *und*

$$\|A\|_w = \max_{i=1}^{n} \frac{1}{w_i} \sum_{j=1}^{n} |a_{i,j}| w_j < 1 \,. \tag{6.2.10}$$

Wir behaupten:

(i) *Es existiert genau ein* $\hat{x} \in \mathbb{K}^n$ *mit*

$$A\hat{x} + b = \hat{x} \,. \tag{6.2.11}$$

(ii) *Für alle* $x_0 \in \mathbb{K}^n$ *strebt die durch die Vorschrift*

$$x_{m+1} = Ax_m + b \quad (m \in \mathbb{N}) \tag{6.2.12}$$

definierte Folge $(x_m)_0^\infty$ *für* $m \to \infty$ *gegen* $\hat{x}$.

(iii) *Es gelten die a priori Abschätzung*

$$\|x_m - \hat{x}\|_w \leqq \frac{\|A\|_w^m}{1 - \|A\|_w} \|x_1 - x_0\|_w \quad (m \in \mathbb{N}) \tag{6.2.13}$$

und die a posteriori Abschätzung

$$\|x_m - \hat{x}\|_w = \frac{\|A\|_w}{1 - \|A\|_w} \|x_m - x_{m-1}\|_w \quad (m \in \mathbb{N}, \neq 0) \,. \tag{6.2.14}$$

Zum *Beweis* beachtet man, daß $\|A\|_w$ die Operatornorm von A im Banach-Raum $(\mathbb{K}^n, \|\ \|_w)$ ist.

Angemerkt sei, daß man den Satz (6.2.9) ebenso bezüglich einer beliebigen anderen Matrixnorm im $M(n \times n, \mathbb{K})$ hätte formulieren können. Zur Herleitung einer normunabhängigen Bedingung an A beweisen wir den folgenden

(6.2.15) **Hilfssatz.** *Es sei* $A \in M(n \times n, \mathbb{C})$. *Dann ist*

$$\rho(A) = \inf \{\|A\| : \|\ \| \ \textit{Matrixnorm in}\ M(n \times n, \mathbb{C})\} \,, \tag{6.2.16}$$

d.h. es wird das Infimum über alle Matrixnormen in $M(n \times n, \mathbb{C})$ *gebildet. Weiter hat man für jede feste Matrixnorm*

$$\rho(A) = \lim_{m \to \infty} \sqrt[m]{\|A^m\|} \,. \tag{6.2.17}$$

Beweis. Nach der Bemerkung (3.3.11), (iii) gilt für jede beliebige Matrixnorm

$$\rho(A) \leqq \|A\| .$$

Zu (6.2.16) bleibt daher zu zeigen, daß für jedes $\epsilon > 0$ eine Matrixnorm $\| \|_\epsilon$ existiert, so daß

$$\rho(A) \geqq \|A\|_\epsilon - \epsilon$$

wird. Hierzu geben wir uns ein $\epsilon > 0$ vor. Bekanntlich gibt es eine invertierbare Matrix $S \in M(n \times n, \mathbb{C})$, so daß $J = S^{-1}AS$ in Jordanscher Normalform

$$J = \begin{pmatrix} D_1 & & 0 \\ & D_2 & \\ & \ddots & \\ 0 & & D_k \end{pmatrix}$$

ist; darin haben die D_i die Gestalt

$$D_i = \begin{pmatrix} \lambda_i & 1 & & 0 \\ & \lambda_i & 1 & \\ & & \ddots & 1 \\ 0 & & & \lambda_i \end{pmatrix} \qquad (i = 1, \ldots, k) ,$$

und die λ_i sind die (nicht notwendig verschiedenen) Eigenwerte von A. Mittels der "Shearing"-Transformation

$$T(\epsilon) = \begin{pmatrix} 1 & & & 0 \\ & \epsilon & & \\ & & \epsilon^2 & \\ 0 & & & \epsilon^{n-1} \end{pmatrix} , \quad J(\epsilon) = T(\epsilon)^{-1} J T(\epsilon)$$

bringen wir J weiter auf die modifizierte Jordansche Normalform

$$J(\epsilon) = \begin{pmatrix} D_1(\epsilon) & & 0 \\ & D_2(\epsilon) & \\ & \ddots & \\ 0 & & D_k(\epsilon) \end{pmatrix}$$

mit

$$D_i(\epsilon) = \begin{pmatrix} \lambda_i & \epsilon & & 0 \\ & \lambda_i & \epsilon & \\ & & \ddots & \epsilon \\ 0 & & & \lambda_i \end{pmatrix} \qquad (i = 1, \ldots, k) .$$

Wir setzen $Q(\epsilon) = S\,T(\epsilon)$; diese Matrix ist offensichtlich invertierbar, so daß wir mit ihr durch

$$\|B\|_\epsilon := \|Q(\epsilon)^{-1} B\, Q(\epsilon)\|_\infty \qquad (B \in M(n \times n, \mathbb{C}))$$

eine Matrixnorm $\|\ \|_\epsilon$ in $M(n \times n, \mathbb{C})$ definieren können. Hierfür gilt dann

$$\|A\|_\epsilon = \|J(\epsilon)\|_\infty \leqq \max_{i=1}^{n} |\lambda_i| + \epsilon = \rho(A) + \epsilon ,$$

womit (6.2.16) bewiesen ist.

Nun zu (6.2.17): Der Beziehung

$$S^{-1} A^m S = J^m = \begin{pmatrix} \lambda_1^m & & & \\ & \ddots & & \\ & & \ddots & \\ 0 & & & \lambda_k^m \end{pmatrix}$$

entnimmt man, daß $\mu \in \mathbb{C}$ genau dann Eigenwert von A^m ist, wenn es einen Eigenwert λ von A mit $\mu = \lambda^m$ gibt. Dies impliziert

$$\rho(A)^m = \rho(A^m)$$

und daher nach (6.2.16) für jede beliebige Matrixnorm

$$(6.2.18) \qquad \rho(A) \leqq \sqrt[m]{\|A^m\|} .$$

Weiter erhält man unter Benutzung der Beziehung

$$A^m = (Q(\epsilon)\, J(\epsilon)\, Q(\epsilon)^{-1})^m = Q(\epsilon)\, (J(\epsilon))^m\, Q(\epsilon)^{-1}$$

mit

$$c(\epsilon) := \|Q(\epsilon)\|_\infty \, \|Q(\epsilon)^{-1}\|_\infty$$

die Abschätzung

$$(6.2.19) \qquad \|A^m\|_\infty \leqq c(\epsilon)\, \|J(\epsilon)\|_\infty^m \leqq c(\epsilon)\, (\rho(A) + \epsilon)^m .$$

Ist $\|\ \|$ eine beliebige Matrixnorm in $M(n \times n, \mathbb{C})$, so gilt, da $M(n \times n, \mathbb{C})$ ein endlichdimensionaler Vektorraum ist, nach (3.3.3) mit einem geeigneten $\gamma > 0$ für $B \in M(n \times n, \mathbb{C})$

$$\|B\| \leqq \gamma\, \|B\|_\infty .$$

Insgesamt ergibt sich infolgedessen nach (6.2.18), (6.2.19)

$$\rho(A) \leqq \sqrt[m]{\|A^m\|} \leqq \sqrt[m]{c(\epsilon)\gamma}\,(\rho(A)+\epsilon)\,;$$

somit ist wegen

$$\lim_{m\to\infty} \sqrt[m]{c(\epsilon)\gamma} = 1$$

und $\epsilon > 0$ beliebig auch die Aussage (6.2.17) bewiesen.

Damit kommen wir zu der folgenden allgemeinen Konvergenzaussage.

(6.2.20) **Satz.** *Es sei* $A \in M(n \times n, \mathbb{C})$, $b \in \mathbb{C}^n$.

Wir behaupten: Das Iterationsverfahren

$$x_{m+1} = Ax_m + b \quad (m \in \mathbb{N})$$

ist genau dann konvergent, wenn $\rho(A) < 1$ *gilt. Ist* $\rho(A) < 1$, $\hat{x}$ *der Grenzwert der Folge* $(x_m)_0^\infty$, *also*

$$\hat{x} = A\hat{x} + b$$

und $|\ |$ *eine Norm im* $\mathbb{C}^n$, *so existiert zu jedem* $k > \rho(A)$ *ein* $N \in \mathbb{N}$, *so daß man für jedes* $m \geqq N$

$$|x_m - \hat{x}| \leqq k^m\,|x_1 - \hat{x}| \tag{6.2.21}$$

hat; ferner gibt es zu jedem $0 \leqq k < \rho(a)$ *Vektoren* x_0, $b \in \mathbb{C}^n$, *so daß* (6.2.21) *für jedes* $m \in \mathbb{N}$ *verletzt ist.*

Zum *Beweis* kann man sowohl (6.2.16) als auch (6.2.17) heranziehen; wir werden uns im Folgenden auf (6.2.17) stützen.

Es sei zunächst $\rho(A) < 1$ vorausgesetzt; weiter sei $|A|$ die Operatornorm von A bezüglich der vorgegebenen Norm $|\ |$. Wegen

$$\lim_{m\to\infty} \sqrt[m]{|A^m|} < 1$$

hat man

$$\sum_{m=0}^{\infty} |A^m| < \infty$$

und daher nach Satz (6.2.1) die Konvergenz des Iterationsverfahrens. Mit dem Operator T aus dem Beweis zu jenem Satz gilt

$$\hat{x} = T^m(\hat{x}),\ x_m = T^m(x_0) \quad (m \in \mathbb{N})$$

und infolgedessen, falls $\rho(A) < k$ ist, mit einem geeigneten $N \in \mathbb{N}$ für alle $m \geqq N$

$$|x_m - \hat{x}| \leqq |T^m|\,|x_0 - \hat{x}| = |A^m|\,|x_0 - x| \leqq k^m\,|x_0 - x|\,.$$

Setzt man $b = 0$, so ergibt sich wegen

$$A\hat{x} = \hat{x}$$

und der Eindeutigkeit von $\hat{x}$ notwendig $\hat{x} = 0$. Ist nun λ Eigenwert von A mit $|\lambda| = \rho(A)$, $x_0 \neq 0$ ein zugehöriger Eigenvektor, so erhält man

$$x_m = T^m(x_0) = A^m x_0 = \lambda^m x_0 ,$$

folglich

$$|x_m - \hat{x}| = |x_m| = |\lambda|^m |x_0| = \rho(A)^m |x_0 - \hat{x}| ,$$

so daß (6.2.21), falls $0 \leqq k < \rho(A)$ ist, bei diesen Werten von x_0 und b für alle $m \in \mathbb{N}$ verletzt ist.

Zu zeigen bleibt, daß im Fall $\rho(A) \geqq 1$ das Iterationsverfahren nicht konvergent ist. Wegen

$$x_m = T^m(x_0) = A^m x_0 + \sum_{\nu=1}^{m-1} A^\nu b \qquad (m \in \mathbb{N})$$

wählen wir $x_0 = 0$, $b = c \neq 0$ Eigenvektor zu λ mit $|\lambda| = \rho(A)$ und erhalten

$$x_{m+1} - x_m = \lambda^m c, \not\to 0 \qquad (m \to \infty) .$$

Damit ist der Beweis des Satzes (6.2.20) abgeschlossen.

Anmerkungen: $\rho(A)$ ist im allgemeinen nicht bekannt, so daß man in der Praxis meist doch auf hinreichende Bedingungen des Typs (6.2.10) angewiesen ist. Selbst wenn man $\rho(A)$ berechnet hätte, ließe sich die Ungleichung (6.2.21) nicht als Fehlerabschätzung verwenden, da man N meist nur schwerlich angeben kann und vor allem $\hat{x}$ nicht bekannt ist. (6.2.21) enthält aber eine Aussage darüber, wie sich die Näherung asymptotisch verbessert; sie sagt nämlich aus, daß man schließlich pro Iterationsschritt ca. $-{}^{10}\log \rho(A)$ Dezimalstellen gewinnt. Aus diesem Grund bezeichnet man

$$r(A) := -\log \rho(A)$$

als das Konvergenzverhältnis des mit A gebildeten Iterationsverfahrens.

Als erste Anwendung von Satz (6.2.9) wollen wir die *Nachiteration* beim Gaußschen Eliminationsverfahren behandeln. Vorgegeben sei ein Gleichungssystem

$$Cx = b$$

mit invertierbarer Matrix $C \in M(n \times n, \mathbb{K})$, $b \in \mathbb{K}^n$. Wir notieren zunächst den

(6.2.22) **Hilfssatz.** *Es sei* $\tilde{C} \in M(n \times n, \mathbb{K})$ *invertierbar mit*

$$(6.2.23) \qquad \|I - \tilde{C}^{-1} C\|_w < 1 .$$

Dann gilt: Für alle $x_0 \in \mathbb{K}^n$ *konvergiert die Folge* $(x_m)_0^\infty$ *mit*

(6.2.24) $\quad \tilde{C}x_{m+1} = (\tilde{C} - C)x_m + b \qquad (m \in \mathbb{N})$

gegen die Lösung $\hat{x}$ *von* $C\hat{x} = b$. *Mit* $A := I - \tilde{C}^{-1}C$ *sind die Ungleichungen* (6.2.13), (6.2.14) *erfüllt.*

Zum *Beweis* beachten wir, daß (6.2.24) der Vorschrift

$$x_{m+1} = (I - \tilde{C}^{-1}C)x_m + \tilde{C}^{-1}b \qquad (m \in \mathbb{N})$$

äquivalent ist. Nach Satz (6.2.9), auf $A = I - \tilde{C}^{-1}C$ angewendet, konvergiert die Folge $(x_m)_0^\infty$ gegen $\hat{x}$ mit

$$(I - \tilde{C}^{-1}C)\hat{x} = \hat{x} - \tilde{C}^{-1}b\,,$$

also, wie behauptet,

$$C\hat{x} = b\,.$$

Die vorgegebene Matrix C werde mit Gauß-Elimination bei halbmaximaler Pivotwahl zerlegt. Gemäß Satz (3.6.6) gewinnen wir als Zerlegungsmatrizen eine normierte untere Dreiecksmatrix $\tilde{L}$ und eine obere Dreiecksmatrix $\tilde{R}$, so daß mit einer Permutationsmatrix P eine Darstellung

$$\tilde{L}\tilde{R} = P(C + F)$$

gilt. Hierzu setzen wir die Abschätzung

(6.2.25) $\quad \|C^{-1}F\|_w < \frac{1}{2}$

voraus. Dann gilt die

(6.2.26) **Bemerkung.** Die Matrix

$$\tilde{C} := C + F$$

erfüllt die Voraussetzung (6.2.23). Die Folge $(x_m)_0^\infty$ in (6.2.24) läßt sich gemäß folgender Vorschrift konstruieren:

$$(6.2.27) \quad \begin{cases} r_m := b - Cx_m\,, \\ \tilde{L}f_m := Pr_m\,, \\ \tilde{R}d_m := f_m\,, \\ x_{m+1} := x_m + d_m\,. \end{cases} \qquad (m = 0, 1, 2, \ldots)$$

Beweis. Zunächst haben wir

$$\tilde{C} = C + F = C(I + C^{-1}F),\ \|C^{-1}F\|_w < \tfrac{1}{2} < 1\,,$$

folglich ist $\tilde{C}$ invertierbar; außerdem gilt

$$(C + F)^{-1} = (I + C^{-1}F)^{-1}C^{-1}$$

und daher

$$I - \tilde{C}^{-1}C = I - (I + C^{-1}F)^{-1} = (I + C^{-1}F)^{-1}C^{-1}F .$$

Es ergibt sich

$$\|I - \tilde{C}^{-1}C\|_w \leqq \|C^{-1}F\|_w \, \|(I + C^{-1}F)^{-1}\|_w < \frac{0{,}5}{1 - 0{,}5} = 1 .$$

Für x_{m+1} in (6.2.27) erhalten wir die Beziehungen

$$\tilde{R}x_{m+1} = \tilde{R}x_m + \tilde{R}d_m = \tilde{R}x_m + f_m$$

und weiter

$$\tilde{L}\tilde{R}x_{m+1} = \tilde{L}\tilde{R}x_m + \tilde{L}f_m = \tilde{L}\tilde{R}x_m + Pr_m = \tilde{L}\tilde{R}x_m + P(b - Cx_m) ,$$

also

$$P\tilde{C}x_{m+1} = P(\tilde{C} - C)x_m + Pb$$

und damit (6.2.24).

In der Übungsaufgabe 6.4 wird gezeigt, daß die Folge $(x_m)_0^\infty$ auch dann gegen $\hat{x}$ konvergiert, wenn bei der Auflösung der Gleichungssysteme $\tilde{L}f_m = Pr_m$, $\tilde{R}d_m = f_m$ (von m abhängige) Rundungsfehler auftreten. Für den Konvergenzbeweis benötigen wir jedoch, daß die r_m und $x_m + d_m$ exakt berechnet werden.

In der praktischen Anwendung wählt man x_0 als die mit Gauß-Elimination gewonnene numerische Lösung von $Cx = b$. Hierzu, insbesondere zur Dreieckszerlegung von C, wird eine einfachgenaue (z.B. REAL*4-) Arithmetik benutzt, ebenso bei der Auflösung der Gleichungssysteme $\tilde{L}f_m = Pr_m$, $\tilde{R}d_m = f_m$ $(m = 0, 1, 2, \ldots)$. Dagegen werden $r_m = b - Cx_m$, $x_{m+1} = x_m + d_m$ doppeltgenau (z.B. mit REAL*8-Arithmetik) berechnet. Mit diesem Verfahren erreicht man im allgemeinen nach wenigen Iterationen eine Näherung für $\hat{x}$, die eine ähnliche Genauigkeit besitzt wie der Wert, den man bei Durchführung der Elimination in doppeltgenauer Arithmetik gewinnt. Wegen der Rundungsfehler in den r_m ist diese Genauigkeit nicht weiter zu verbessern. Das Verfahren der Nachiteration kann Rechenzeitvorteile gegenüber der doppeltgenauen Elimination bringen: dazu beachte man, daß pro Iteration nur n^2 doppeltgenaue Rechenoperationen auftreten.

(6.2.28) **Zahlenbeispiel.** Wir wenden das angegebene Verfahren auf das Gleichungssystem (2.3.3) an, wobei wir die in Beispiel (1.1.16) erwähnten REAL*4 und REAL*8-Arithmetiken benutzen. Wir erhalten, auf insgesamt 12 Stellen gerundet, folgende Näherungen:

m	x_m			
0	−20,7628021240	−2,74793243408	14,7451000214	2,61587429047
1	−20,7627241762	−2,74791920243	14,7450338356	2,61586313795
2	−20,7627241768	−2,74791920250	14,7450338360	2,61586313801

Die Residuen r_m (auf 8 bzw. 7 Stellen gerundet) lauten:

m	r_m			
0	$-0{,}95500946 \cdot 10^{-4}$	$-0{,}29659271 \cdot 10^{-4}$	$-0{,}62847137 \cdot 10^{-4}$	$-0{,}18024445 \cdot 10^{-4}$
1	$0{,}1683986 \cdot 10^{-11}$	$0{,}4236256 \cdot 10^{-10}$	$-0{,}5756107 \cdot 10^{-10}$	$-0{,}4273559 \cdot 10^{-10}$
2	$-0{,}1065814 \cdot 10^{-13}$	$-0{,}3552714 \cdot 10^{-14}$	$0{,}3907985 \cdot 10^{-13}$	$0{,}1776357 \cdot 10^{-13}$

Die folgenden Residuen besitzen die gleiche Größenordnung wie r_2; entsprechend wird die Näherung x_2 nicht weiter verbessert. Die Lösung des Gleichungssystem mit doppeltgenauer Elimination liefert ein Ergebnis, das in den angegebenen Dezimalen mit x_2 übereinstimmt.

6.3. Das Gesamtschritt- und das Einzelschrittverfahren

Wir beschäftigen uns in diesem Abschnitt mit der iterativen Behandlung linearer Gleichungssysteme der Gestalt

$$(6.3.1) \qquad \begin{cases} Gx = h\,; \\ G \in M(n \times n, \mathbb{C}),\ = (g_{i,j})_{(n,n)};\ h \in \mathbb{C}^n,\ = (\gamma_i)_1^n\,. \end{cases}$$

Für die Anwendung der in Abschnitt 6.2 bewiesenen Sätze werden geeignete Umformungen durchgeführt. Hierzu teilt man die Matrix $G =: D - C_1 - C_2$ auf; dabei ist

D eine Diagonalmatrix, $= \operatorname{diag}(g_{1,1}, \ldots, g_{n,n})$,
C_1 eine untere, C_2 eine obere } Dreiecksmatrix mit Nullen in der Diagonalen.

Symbolisch schreibt man hierfür kurz

$$G = \begin{pmatrix} & & -C_2 \\ & D & \\ -C_1 & & \end{pmatrix}\,.$$

Für das Folgende setzen wir voraus

$$(6.3.2) \qquad D \text{ invertierbar, d.h. } \forall i \in \{1, 2, \ldots, n\}\ g_{i,i} \neq 0\,.$$

Offenbar ist die Gleichung (6.3.1) äquivalent umformbar in

$$(6.3.3) \qquad D^{-1}(C_1 + C_2)x + D^{-1}h = x\,.$$

Mit

$$A = A_{Ges} := D^{-1}(C_1 + C_2),\ b := D^{-1}h$$

ist so (6.3.1) auf ein im Abschnitt 6.2 behandeltes Problem zurückgeführt. Die dortige Iterationsvorschrift lautet, wenn man

$$x_m = \begin{pmatrix} \xi_m^1 \\ \vdots \\ \xi_m^n \end{pmatrix}$$

setzt, komponentenweise

$$\xi_{m+1}^i = \frac{1}{g_{i,i}} \left\{ \sum_{\substack{j=1 \\ j \neq i}}^{n} (-g_{i,j}) \xi_m^j + \gamma_i \right\} \quad (i = 1, \ldots, n;\ m = 0, 1, 2, \ldots) \tag{6.3.4}$$

Das hierdurch beschriebene Iterationsverfahren nennt man Gesamtschritt- oder auch Jacobi-Verfahren. Hinreichende Bedingungen für die Konvergenz werden unten angegeben.

Mit D ist auch die Matrix $(D - C_1)$ invertierbar. Wir können daher die Gleichung (6.3.1) auch in die Form

$$(D - C_1)^{-1} C_2 x + (D - C_1)^{-1} h = x \tag{6.3.5}$$

bringen; folglich ist (6.3.1) ebenfalls mit

$$A = A_{Ein} := (D - C_1)^{-1} C_2,\ b := (D - C_1)^{-1} h$$

einem Problem des in Abschnitt 6.2 betrachteten Typs äquivalent. Die dortige Iterationsvorschrift ergibt hier komponentenweise

$$\left\{ \xi_{m+1}^i = \frac{1}{g_{i,i}} \left\{ \sum_{j=1}^{i-1} (-g_{i,j}) \xi_{m+1}^j + \sum_{j=i+1}^{n} (-g_{i,j}) \xi_m^j + \gamma_i \right\} \right.$$
$$\left. (i = 1, \ldots, n;\ m = 0, 1, 2, \ldots) \right. . \tag{6.3.6}$$

Dieses Verfahren bezeichnet man als das Einzelschritt- oder auch Gauß-Seidel-Verfahren. Zur Berechnung der Komponenten von x_{m+1} zieht man hier im Gegensatz zum Gesamtschrittverfahren die bereits berechneten Komponenten so weit wie möglich heran.

Nun zur Konvergenz der beiden Verfahren: Offenbar ist mit $w \in \mathbb{R}^n, > 0$ und $0 \leq p < 1$ die Bedingung $\|A\|_w \leq p$ genau dann erfüllt, wenn – vgl. Definition (3.5.5) –

$$\hat{A} w \leq p w$$

gilt. Im Fall des Gesamtschrittverfahrens $A = A_{Ges}$ erschließt man auf Grund der speziellen Gestalt der Matrizen D, C_1 und C_2 die Beziehung

$$\hat{A} = \hat{D}^{-1} \widehat{(C_1 + C_2)} = \hat{D}^{-1} (\hat{C}_1 + \hat{C}_2) .$$

Hiermit folgt sofort, daß $\|A\|_w \leqq p$ genau dann zutrifft, wenn

(6.3.7) $\quad (\hat{C}_1 + \hat{C}_2)w \leqq p\hat{D}w$

abschätzbar ist. Im Fall des Einzelschrittverfahrens hat man trivialerweise

$$\hat{A}_{Ein} \leqq \widehat{(D - C_1)^{-1}}\hat{C}_2 .$$

Da $D^{-1}C_1$ als echte untere Dreiecksmatrix nilpotent ist, erhält man

$$\begin{aligned}(D - C_1)^{-1} &= (I - D^{-1}C_1)^{-1}D^{-1} \\ &= \{I + (D^{-1}C_1) + (D^{-1}C_1)^2 + \ldots + (D^{-1}C_1)^{n-1}\}D^{-1},\end{aligned}$$

mithin

$$\begin{aligned}\widehat{(D - C_1)^{-1}} &\leqq \{I + (\hat{D}^{-1}\hat{C}_1) + (\hat{D}^{-1}\hat{C}_1)^2 + \ldots + (\hat{D}^{-1}\hat{C}_1)^{n-1}\}\hat{D}^{-1} \\ &= (\hat{D} - \hat{C}_1)^{-1} .\end{aligned}$$

Daher ergibt sich, daß $\|A_{Ein}\|_w \leqq p$ sicherlich dann gilt, wenn

(6.3.8) $\quad (\hat{D} - \hat{C}_1)^{-1}\hat{C}_2 w \leqq pw$

bzw. schärfer – man beachte $(\hat{D} - \hat{C}_1)^{-1} \geqq 0$ –

(6.3.8′) $\quad \hat{C}_2 w \leqq p(\hat{D} - \hat{C}_1)w$

eintritt.

Zusammenfassend notieren wir den

(6.3.9) **Satz.** *Hinreichend für die Konvergenz des Einzelschrittverfahrens* (6.3.6) *ist die Existenz eines Vektors* $w \in \mathbb{R}^n, > 0$ *und einer reellen Zahl* $0 \leqq p < 1$ *mit* (6.3.8) *bzw. schärfer* (6.3.8′). *Ist darüberhinaus sogar die Bedingung* (6.3.7) *erfüllt, so ist neben dem Einzelschrittverfahren auch das Gesamtschrittverfahren* (6.3.4) *konvergent. Der Grenzwert* $\hat{x}$ *der durch das Gesamt- bzw. Einzelschrittverfahren definierten Folge* $(x_m)_0^\infty$ *genügt der Gleichung* (6.3.1). *Abschätzungen des Abbruchfehlers sind durch*

$$(6.3.10) \quad \|x_m - \hat{x}\|_w \leqq \frac{p^n}{1-p}\|x_1 - x_0\|_w \qquad (m \in \mathbb{N})$$

bzw.

$$(6.3.11) \quad \|x_m - \hat{x}\|_w \leqq \frac{p}{1-p}\|x_m - x_{m-1}\|_w \quad (m \in \mathbb{N}, \neq 0)$$

gegeben.

Dieser Satz sagt unter anderem aus, daß die Ungleichung (6.3.7), die hinreichend für die Konvergenz des Gesamtschrittverfahrens ist, die Bedingung (6.3.8′), die zur Konvergenz des Einzelschrittverfahrens führt, impliziert. Das bedeutet jedoch *nicht*, daß die Konvergenz des Gesamtschrittverfahrens stets die Konvergenz des Einzelschrittverfahrens nach sich zieht. Vergleiche hierzu die Übungsaufgabe 6.5!

Wir notieren als nahezu unmittelbare Folgerung den

(6.3.12) **Satz.** *Erfüllt die Matrix* G *das starke Zeilensummenkriterium*

$$(6.3.13)\qquad \sum_{\substack{j=1\\ j\neq i}}^{n} |g_{i,j}| < |g_{i,i}| \qquad (i = 1, \ldots, n)$$

oder das starke Spaltensummenkriterium

$$(6.3.14)\qquad \sum_{\substack{i=1\\ i\neq j}}^{n} |g_{i,j}| < |g_{j,j}| \qquad (j = 1, \ldots, n)\,,$$

so ist sowohl das Gesamtschrittverfahren als auch das Einzelschrittverfahren konvergent.

Beweis.

a) Es genüge G zunächst dem starken Zeilensummenkriterium. Setzt man dann

$$p := \max_{i=1}^{n} \frac{1}{|g_{i,i}|} \sum_{\substack{j=1\\ j\neq i}}^{n} |g_{i,j}|\,,$$

so gilt $0 \leqq p < 1$ und weiter mit $w = (1)_1^n$ die Bedingung (6.3.7).

b) Erfüllt G das starke Spaltensummenkriterium, so genügt die Matrix

$$\tilde{G} = \begin{pmatrix} g_{n,n} & \cdots & g_{1,n} \\ \vdots & & \vdots \\ g_{n,1} & \cdots & g_{1,1} \end{pmatrix}$$

dem starken Zeilensummenkriterium. Teilt man

$$\tilde{G} = \begin{pmatrix} \tilde{D} & -\tilde{C}_2 \\ -\tilde{C}_1 & \end{pmatrix}$$

analog G auf, so folgt mit

$$\tilde{A}_{Ges} := \tilde{D}^{-1}(\tilde{C}_1 + \tilde{C}_2),\ \tilde{A}_{Ein} := (\tilde{D} - \tilde{C}_1)^{-1}\tilde{C}_2$$

aus Satz (6.2.20) und a)

$$(6.3.15)\qquad \rho(\tilde{A}_{Ges}) < 1\,,\ \rho(\tilde{A}_{Ein}) < 1\,.$$

Die Permutationsmatrix

$$P = \begin{pmatrix} 0 & \cdots & \cdots & 0 & 1 \\ \vdots & & & & 0 \\ \vdots & & & & \vdots \\ 0 & & & & \vdots \\ 1 & 0 & \cdots & \cdots & 0 \end{pmatrix}$$

hat die Eigenschaft $P^t = P$, $P^{-1} = P$; ferner gilt

$$\tilde{D} = P^{-1} D^t P, \quad \tilde{C}_1 = P^{-1} C_1^t P, \quad \tilde{C}_2 = P^{-1} C_2^t P .$$

Damit erhält man nach (3.3.11)

$$\begin{aligned} \rho(\tilde{A}_{Ges}) &= \rho(\tilde{D}^{-1}(\tilde{C}_1 + \tilde{C}_2)) = \rho(P^{-1} D^{t^{-1}} (C_1^t + C_2^t) P) \\ &= \rho(D^{t^{-1}} (C_1^t + C_2^t)) = \rho((C_1 + C_2) D^{-1}) \\ &= \rho(D^{-1}(C_1 + C_2) D^{-1} D) = \rho(A_{Ges}) \end{aligned}$$

sowie nahezu völlig analog

$$\rho(\tilde{A}_{Ein}) = \rho(A_{Ein}) ,$$

womit unter Berücksichtigung von (6.3.16) abermals nach Satz (6.2.20) auch die Behauptung bezüglich des Spaltenkriteriums bewiesen ist.

Die Voraussetzung des starken Zeilen- bzw. Spaltensummenkriteriums ist bei Anwendungen, insbesondere bei der numerischen Behandlung von Randwertproblemen partieller Differentialgleichungen i.a. nicht gegeben. Eine deutlich schwächere Voraussetzung, die die Konvergenz zumindest noch des Einzelschrittverfahrens nach sich zieht, enthält der

(6.3.16) **Satz.** *Es sei* $w = (1)_1^n$, $v \in \mathbb{R}^n$, ≥ 0 *die (eindeutig existierende) Lösung der Gleichung*

$$(\hat{D} - \hat{C}_1) v = \hat{C}_2 w \tag{6.3.17}$$

und $p := \|v\|_\infty$.

Dann ist bezüglich w *und* p *die Ungleichung* (6.3.8) *erfüllt, mithin das Einzelschrittverfahren konvergent, sofern* $p < 1$ *gilt.*

Der *Beweis* ist sofort erbracht: Wegen $v \leqq pw$ kann man nach (6.3.17)

$$(\hat{D} - \hat{C}_1)^{-1} \hat{C}_2 w = v \leqq pw$$

abschätzen.

Die Konvergenz des Einzel- und Gesamtschrittverfahrens garantiert in Verschärfung des Kriteriums (6.3.11) der

(6.3.18) **Satz.** *Erfüllt die Matrix* G *das schwache Zeilensummenkriterium*

$$(6.3.19)\quad \begin{cases} \sum\limits_{\substack{j=1\\ j\neq i}}^{n} |g_{i,j}| \leqq |g_{i,i}| & (i = 1, \ldots, n)\,. \\ \sum\limits_{j=i+1}^{n} |g_{i,j}| < |g_{i,i}| & (i = 1, \ldots, n-1) \end{cases}$$

oder das schwache Spaltensummenkriterium

$$(6.3.20)\quad \begin{cases} \sum\limits_{\substack{i=1\\ i\neq j}}^{n} |g_{i,j}| \leqq |g_{j,j}| & (j = 1, \ldots, n)\,. \\ \sum\limits_{i=1}^{j-1} |g_{i,j}| < |g_{j,j}| & (j = 2, \ldots, n)\,, \end{cases}$$

so ist sowohl das Einzel- als auch das Gesamtschrittverfahren konvergent.

Beweis. Die Matrix G genüge dem schwachen Zeilensummenkriterium. Wir beweisen zunächst die Konvergenz des Einzelschrittverfahrens; hierzu stützen wir uns auf den vorangehenden Satz (6.3.16). Zu zeigen ist, daß $p = \|v\|_\infty < 1$ ist, d.h. daß sämtliche Komponenten v_l $(\geqq 0)$ kleiner als 1 sind. Wir beweisen dies durch Induktion nach l: Für $l = 1$ gilt nach (6.3.17)

$$|g_{1,1}|\, v_1 = \sum_{j=2}^{n} |g_{1,j}|\,,$$

also gemäß der Voraussetzung (6.3.19)

$$v_1 = \frac{1}{|g_{1,1}|} \sum_{j=2}^{n} |g_{1,j}| < 1\,.$$

Wir nehmen nun an, es sei bereits für $i = 1, \ldots, l-1$ $v_i < 1$ nachgewiesen. Falls $l \leqq n$ ist, hat man

$$|g_{l,l}|\, v_l = \sum_{j=1}^{l-1} |g_{l,j}|\, v_j + \sum_{j=l+1}^{n} |g_{l,j}|\,.$$

Ist nun

$$\sum_{\substack{j=1\\ j\neq l}}^{n} |g_{l,j}| = |g_{l,l}|\,,$$

so muß notwendig $\sum_{j=1}^{l-1} |g_{l,j}| > 0$ sein, woraus unter Benutzung der Induktionsannahme sofort

$$v_l < \frac{1}{|g_{l,l}|} \sum_{\substack{j=1 \\ j \neq l}}^{n} |g_{l,j}| = 1$$

folgt. Gilt sogar

$$\sum_{\substack{j=1 \\ j \neq l}}^{n} |g_{l,j}| < |g_{l,l}| ,$$

so erschließt man hier mit Hilfe der Induktionsvoraussetzung

$$v_l \leqq \frac{1}{|g_{l,l}|} \sum_{\substack{j=1 \\ j \neq l}}^{n} |g_{l,j}| < 1 ,$$

womit die Induktionsbehauptung gänzlich, also $p < 1$ bewiesen ist.

Nun zur Konvergenz des Gesamtschrittverfahrens: Es sei $A = A_{Ges}$. Gemäß (6.3.19) hat man $\|\hat{A}\|_\infty \leqq 1$, folglich für $m \in \mathbb{N}$ $\|\hat{A}^m\|_\infty \leqq 1$. Für $x = (\xi^l)_1^n \in \mathbb{R}^n$ setzen wir

$$x_m = (\xi_m^l)_{l=1}^n := \hat{A}^m x$$

und schätzen

$$\|x_m\| \leqq \|\hat{A}^m\|_\infty \, \|x\|_\infty \leqq \|x\|_\infty \tag{6.3.21}$$

ab. Weiter beweisen wir mit den v_l aus dem vorangehenden Konvergenzbeweis des Einzelschrittverfahrens durch Induktion nach m:

$$|\xi_m^l| \leqq v_l \, \|x\|_\infty \qquad (l = 1, \ldots, m;\ m = 1, \ldots, n) . \tag{6.3.22}$$

Für $m = 1$ ist – vgl. (6.3.4) –

$$|\xi_1^1| \leqq \frac{1}{|g_{1,1}|} \sum_{j=2}^{n} |g_{1,j}| \, |\xi^j| \leqq v_1 \, \|x\|_\infty .$$

Wir nehmen nun an, (6.3.22) gelte für $m - 1$, und es sei $m \leqq n$. Dann ergibt sich unter Berücksichtigung von (6.3.21) für $l = 1, \ldots, m$

$$\begin{aligned} |\xi_m^l| &\leqq \frac{1}{|g_{l,l}|} \sum_{\substack{j=1 \\ j \neq l}}^{n} |g_{l,j}| \, |\xi_{m-1}^j| \\ &\leqq \frac{1}{|g_{l,l}|} \left\{ \sum_{j=1}^{l-1} |g_{l,j}| \, v_j + \sum_{j=l+1}^{n} |g_{l,j}| \right\} \|x\|_\infty = v_l \, \|x\|_\infty . \end{aligned}$$

Mit $v = (v_i)_1^n$, $p = \|v\|_\infty$ (< 1) liefert (6.3.22) speziell für $m = n$

$$\|\hat{A}^n x\|_\infty \leqq p \|x\|_\infty ,$$

mithin

(6.3.23) $$\|\hat{A}^n\|_\infty \leqq p < 1 .$$

Wegen $\|A^n\|_\infty = \|\hat{A^n}\|_\infty \leqq \|\hat{A}^n\|_\infty$ ist damit nach Satz (6.2.1) unter Berücksichtigung des Hilfssatzes (6.1.3) die Konvergenz des Gesamtschrittverfahrens gezeigt. Abschätzungen des Abbrechfehlers sind bezüglich $\| \|_\infty$ durch (6.2.5), (6.2.6) sowie (6.2.7), (6.2.8) gegeben.

Übrigens kann man die Konvergenz auch mit Hilfe des Satzes (6.2.9) oder (6.3.9) erschließen. Mit $e = (1)_1^n$ folgt nämlich aus (6.3.23) $\hat{A}^n e < e$. Setzt man nun

$$w = (w_i)_1^n := e + \hat{A}e + \ldots + \hat{A}^{n-1}e \quad (> 0) ,$$

so erhält man $\hat{A}w < w$ und dann mit

$$p = \max_{i=1}^{n} \frac{1}{|g_{i,i}|\, w_i} \sum_{\substack{j=1 \\ j \neq i}}^{n} |g_{i,j}|\, w_j \quad (< 1)$$

die Ungleichungen

$$\|A\|_w \leqq p, \quad (\hat{C}_1 + \hat{C}_2)w \leqq p\hat{D}w .$$

Der Beweis im Fall des schwachen Spaltensummenkriteriums wird wie beim Satz (6.3.12) durch Reduktion auf das entsprechende Zeilensummenkriterium geführt.

6.4. Relaxationsverfahren

Wir behandeln hier nochmals das Problem (6.3.1); hierzu übernehmen wir die Bezeichnungen und Voraussetzungen des vorangehenden Abschnitts 6.3.

Wir berechnen für $m \in \mathbb{N}$ aus x_m zunächst

$$D\tilde{x}_{m+1} := C_1 x_m + C_2 x_m + h ,$$

d.h. $\tilde{x}_{m+1}$ aus x_m gemäß der Rechenvorschrift des Gesamtschrittverfahrens. Danach bestimmen wir x_{m+1} durch eine gewisse Mittelung aus x_m und $\tilde{x}_{m+1}$; wir setzen nämlich mit $\omega > 0$

$$x_{m+1} = (1 - \omega)x_m + \omega\tilde{x}_{m+1} .$$

Insgesamt führt dies zu der Rekursionsvorschrift

(6.4.1) $$Dx_{m+1} = \{(1 - \omega)D + \omega(C_1 + C_2)\}x_m + h \quad (m \in \mathbb{N}) .$$

Mit den Abkürzungen

$$A_{Ges}(\omega) := D^{-1}\{(1-\omega)D + \omega(C_1 + C_2)\},\quad b := D^{-1}h \tag{6.4.2}$$

ist (6.3.1) offenbar dem „Fixpunktproblem"

$$A_{Ges}(\omega)x + b = x$$

äquivalent; das hierzu gehörende Iterationsverfahren (6.4.1) nennt man das Relaxationsverfahren in Gesamtschritten. Komponentenweise erhält man

$$\left\{\begin{array}{l} \xi^i_{m+1} = (1-\omega)\xi^i_m + \dfrac{\omega}{g_{i,i}}\left\{\displaystyle\sum_{\substack{j=1\\ j\neq i}}^{n}(-g_{i,j})\xi^j_m + \gamma_i\right\} \\ \\ (i = 1,\ldots,n;\ m = 0,1,2,\ldots)\,. \end{array}\right. \tag{6.4.3}$$

Für $\omega = 1$ geht dieses Verfahren in das Gesamtschrittverfahren über.

Erfüllt $A_{Ges}(\omega)$ bezüglich einer Norm $\|\ \|$ die Konvergenzbedingung (6.2.2), so ist natürlich das Verfahren (6.4.1) konvergent, und es gilt nach (6.2.7) die Fehlerabschätzung

$$\|x_m - \hat{x}\| \leqq \|G^{-1}\|\,\left\|\left(\frac{1}{\omega} - 1\right)D + C_1 + C_2\right\|\,\|x_m - x_{m-1}\|\,. \tag{6.4.4}$$

Diese Abschätzung ist dann praktikabel, wenn man $\|G^{-1}\|$ bestimmen kann, ohne G^{-1} berechnen zu müssen.

Durch ähnliche Überlegungen wie oben gelangt man auch zu dem Relaxationsverfahren in Einzelschritten

$$(D - \omega C_1)x_{m+1} = \{(1-\omega)D + \omega C_2\}x_m + \omega h \quad (m \in \mathbb{N})\,; \tag{6.4.5}$$

dieses lautet komponentenweise

$$\left\{\begin{array}{l} \xi^i_{m+1} = (1-\omega)\xi^i_m + \dfrac{\omega}{g_{i,i}}\left\{\displaystyle\sum_{j=1}^{i-1}(-g_{i,j})\xi^j_{m+1} + \sum_{j=i+1}^{n}(-g_{i,j})\xi^j_m + \gamma_i\right\} \\ \\ (i = 1,\ldots,n;\ m = 0,1,2,\ldots)\,. \end{array}\right. \tag{6.4.6}$$

Für $\omega = 1$ ist dies natürlich das Einzelschrittverfahren. Entsprechend (6.4.2) setzt man hier

$$A_{Ein}(\omega) := (D - \omega C_1)^{-1}\{(1-\omega)D + \omega C_2\};\quad b := \omega(D - \omega C_1)^{-1}h \tag{6.4.7}$$

und stellt fest, daß (6.3.1) in das Problem

$$A_{Ein}(\omega)x + b = x$$

äquivalent umformbar ist. Analog zu (6.4.4) gewinnt man aus (6.2.7) für (6.4.5) die Fehlerabschätzung

$$\|x_m - \hat{x}\| \leqq \|G^{-1}\|\,\left\|\left(\frac{1}{\omega} - 1\right)D + C_2\right\|\,\|x_m - x_{m-1}\|\,. \tag{6.4.8}$$

Um eine möglichst rasche Konvergenz der Iterationsverfahren (6.4.1) und (6.4.5) zu gewährleisten, hat man gemäß Satz (6.2.20) ω so zu wählen, daß der Spektralradius der zugehörigen Iterationsmatrix $A_{Ges}(\omega)$ bzw. $A_{Ein}(\omega)$ minimal wird. Es ist unser weiteres Ziel, einen derartigen „optimalen" Wert ω_{opt} unter geeignete Voraussetzungen an die Matrix G anzugeben; dabei werden wir uns auf das Studium des Relaxationsverfahrens in Einzelschritten beschränken.

Für diese Untersuchungen sind die abkürzenden Bezeichnungen

$$A(\omega) := A_{Ein}(\omega), \quad \tilde{C}_1 := D^{-1}C_1, \quad \tilde{C}_2 := D^{-1}C_2$$

nützlich. Hiermit erhält man für $A(\omega)$ die Darstellung

$$A(\omega) = (I - \omega\tilde{C}_1)^{-1}\{(1-\omega)I + \omega\tilde{C}_2\}, \tag{6.4.9}$$

worin wie bisher I die (n,n)-Einheitsmatrix bezeichnet.

(6.4.10) **Satz.** *Für alle* $\omega > 0$ *gilt*

$$\rho(A(\omega)) \geqq |\omega - 1| .$$

Beweis. Es bezeichne $Q(\mu)$ das charakteristische Polynom der Matrix $A(\omega)$, d.h.

$$Q(\mu) := \det(\mu I - A(\omega)) .$$

Wegen $\det(I - \omega\tilde{C}_1) = 1$ ergibt sich

$$Q(\mu) = \det((I - \omega\tilde{C}_1)(\mu I - A(\omega)))$$

und daraus mit (6.4.9)

$$Q(\mu) = \det((\mu + \omega - 1)I - \mu\omega\tilde{C}_1 - \omega\tilde{C}_2) . \tag{6.4.11}$$

Weiter seien $\lambda_1(\omega), \ldots, \lambda_n(\omega)$ die (nicht notwendig verschiedenen) Eigenwerte der Matrix $A(\omega)$. Unter Beachtung der Beziehung

$$\prod_{i=1}^{n}(-\lambda_i(\omega)) = Q(0) = \det((\omega - 1)I - \omega\tilde{C}_2) = (\omega - 1)^n$$

erschließt man

$$\rho(A(\omega)) = \max_{i=1}^{n} |\lambda_i(\omega)| \geqq |\omega - 1| ,$$

was zu zeigen war.

Der vorstehende Satz besagt, daß für die Konvergenz des Relaxationsverfahrens (6.4.8) $0 < \omega < 2$ notwendig ist. Daher werden wir im Weiteren bei Fragen, die die Konvergenz betreffen, nur derartige Werte des Parameters ω in Betracht zu ziehen haben. Im Fall $0 < \omega < 1$ spricht man von einem Verfahren der Unterrelaxation, im Fall $1 < \omega < 2$ von einem Verfahren der Überrelaxation.

Wir führen nun zunächst den Typ von Matrizen ein, für den wir ω_{opt} bestimmen wollen. Hierzu sei wie bisher $G = D(I - \tilde{C}_1 - \tilde{C}_2)$ zerlegt; weiter bezeichne für $\alpha \in \mathbb{C}, \neq 0$

$$J(\alpha) := \alpha \tilde{C}_1 + \frac{1}{\alpha} \tilde{C}_2 .$$

(6.4.12) **Definition.** Die Matrix G heißt *konsistent geordnet*, wenn die Eigenwerte der Matrizen $J(\alpha)$ von α unabhängig sind, d.h. für alle $\alpha, \beta \in \mathbb{C}, \neq 0$ und alle $\lambda \in \mathbb{C}$ gilt

λ Eigenwert von $J(\alpha) \iff \lambda$ Eigenwert von $J(\beta)$.

Insbesondere sind die Matrizen G, die die „Eigenschaft *A*" (property *A*) besitzen, konsistent geordnet. Dabei versteht man unter Matrizen mit der Eigenschaft *A* schwach besetzte Blockmatrizen des Typs

$$G = \begin{pmatrix} D_1 & -C_{1,2} & 0 & & \\ -C_{2,1} & D_2 & -C_{2,3} & & 0 \\ 0 & -C_{3,2} & \ddots & \ddots & \\ & & \ddots & & D_p \\ & 0 & & D_p & \end{pmatrix},$$

worin die $D_1, D_2, \ldots, D_p$ Diagonalmatrizen eventuell verschiedener Größe und die $C_{i,j}$ entsprechend Rechteckmatrizen sind. Derartige Matrizen treten, wie wir auch noch aufzeigen werden, bei der numerischen Behandlung partieller Differentialgleichungen auf.

Zum *Beweis*, daß Matrizen mit der Eigenschaft *A* konsistent geordnet sind, berechnet man zunächst mit $\tilde{C}_{i,j} = D_i^{-1} C_{i,j}$

$$\tilde{C}_1 = \begin{pmatrix} 0 & & & & \\ \tilde{C}_{2,1} & 0 & & & 0 \\ & \tilde{C}_{3,2} & \ddots & \ddots & \\ & & \ddots & & 0 \\ & 0 & & 0 & \end{pmatrix}, \quad \tilde{C}_2 = \begin{pmatrix} 0 & \tilde{C}_{1,2} & & & \\ & 0 & \tilde{C}_{2,3} & & 0 \\ & & \ddots & \ddots & \\ & & \ddots & & 0 \\ & 0 & & 0 & \end{pmatrix}$$

und ermittelt anschließend mit Hilfe der „Shearing"-Transformation

$$S(\alpha) = \begin{pmatrix} I_1 & & & \\ & \alpha I_2 & & 0 \\ & & \ddots & \\ & 0 & & \alpha^{p-1} I_p \end{pmatrix}$$

für $\alpha \in \mathbb{C}, \neq 0$ die Beziehung

$$J(\alpha) = S(\alpha)^{-1} J(1) S(\alpha),$$

womit bereits die Behauptung bewiesen ist, da bekanntlich ähnliche Matrizen die gleichen Eigenwerte haben.

Grundlegend für die weitere Diskussion ist der

(6.4.13) **Hilfssatz.** *Ist* G *konsistent geordnet, so gilt*

(i) $\lambda \in \mathbb{C}$ *ist Eigenwert von* $A_{Ges} = \tilde{C}_1 + \tilde{C}_2$ *genau dann, wenn* $-\lambda$ *Eigenwert von* A_{Ges} *ist.*

(ii) $\mu \in \mathbb{C}$ *ist genau dann Eigenwert von* $A(\omega) = A_{Ein}(\omega)$, *wenn es einen Eigenwert* $\lambda \in \mathbb{C}$ *von* A_{Ges} *mit*

$$\mu \lambda^2 \omega^2 = (\mu + \omega - 1)^2 \tag{6.4.14}$$

gibt.

Beweis.

(i) Da G konsistent geordnet ist, ist λ Eigenwert von $A_{Ges} = J(1)$ genau dann, wenn λ Eigenwert von $J(-1) = -A_{Ges}$, d.h. $-\lambda$ Eigenwert von A_{Ges} ist.

(ii) Es bezeichne $P(\lambda)$ das charakteristische Polynom von $A_{Ges} = J(1)$, also

$$P(\lambda) = \det(\lambda I - J(1)), \tag{6.4.15}$$

sowie $Q(\mu)$ wie im Beweis zu Satz (6.4.10) das charakteristische Polynom von $A(\omega)$. Für $\mu \in \mathbb{C}, \neq 0$ hat man nach (6.4.11)

$$Q(\mu) = \det\left\{(\mu + \omega - 1) I - \omega \sqrt{\mu} \left(\sqrt{\mu}\, \tilde{C}_1 + \frac{1}{\sqrt{\mu}} \tilde{C}_2\right)\right\}$$

und daher

$$Q(\mu) = (\omega \sqrt{\mu})^n \cdot \det\left(\frac{\mu + \omega - 1}{\omega \sqrt{\mu}} I - J(\sqrt{\mu})\right). \tag{6.4.16}$$

Weiter sei nun zunächst $\mu \in \mathbb{C}$ Eigenwert von $A(\omega)$. Ist $\mu \neq 0$, so ist gemäß

(6.4.16) $\frac{\mu + \omega - 1}{\omega \sqrt{\mu}}$ Eigenwert von $J(\sqrt{\mu})$, folglich, da G konsistent geordnet ist,

auch Eigenwert von $J(1)$. Es existiert also ein Eigenwert λ von A_{Ges} mit

$$\lambda = \frac{\mu + \omega - 1}{\omega \sqrt{\mu}},$$

also auch mit (6.4.14). Ist $\mu = 0$, so hat man nach (6.4.11)

$$0 = Q(0) = \det((\omega - 1) I - \omega \tilde{C}_2) = (\omega - 1)^n,$$

mithin $\omega = 1$. In diesem Falle existiert trivialerweise ein Eigenwert λ von A_{Ges}, der (6.4.14) erfüllt.

Umgekehrt habe man $\lambda, \mu \in \mathbb{C}$; dabei sei λ Eigenwert von A_{Ges}, und es gelte (6.4.14). Im Fall $\mu \neq 0$ löst man (6.4.14) nach λ auf und erhält

$$(*) \quad \lambda = \pm \frac{\mu + \omega - 1}{\omega \sqrt{\mu}} .$$

Da nach (i) mit λ auch $-\lambda$ Eigenwert von A_{Ges} ist, kann man o.E. in (*) das positive Vorzeichen als gegeben annehmen. Danach ist $\frac{\mu + \omega - 1}{\omega \sqrt{\mu}}$ Eigenwert von $J(1) = A_{Ges}$, also auch von $J(\sqrt{\mu})$ und somit nach (6.4.16) μ Eigenwert von $A(\omega)$. Im Fall $\mu = 0$ folgt aus (6.4.14) notwendig $\omega = 1$; nach (6.4.11) ergibt sich dann

$$Q(0) = \det \tilde{C}_2 = 0 ,$$

d.h. $\mu = 0$ ist Eigenwert von $A(\omega) = A(1)$.

Aus Hilfssatz (6.4.13) erhalten wir als unmittelbare

(6.4.17) **Folgerung.** Ist G konsistent geordnet, so gilt

$$\rho(A_{Ein}) = (\rho(A_{Ges}))^2 .$$

Ist also G konsistent geordnet, so impliziert die Konvergenz des Gesamtschrittverfahrens die Konvergenz des Einzelschrittverfahrens, wobei die Konvergenzgeschwindigkeit des Einzelschrittverfahrens gegenüber der des Gesamtschrittverfahrens asymptotisch etwa doppelt so groß ist.

Zum *Beweis* der Folgerung (6.4.17) wenden wir den vorangehenden Hilfssatz für $\omega = 1$ an. Danach ist μ Eigenwert von $A_{Ein} = A(1)$ genau dann, wenn zu A_{Ges} ein Eigenwert λ mit $\mu\lambda^2 = \mu^2$ existiert, d.h. wenn $\mu = 0$ ist oder A_{Ges} einen Eigenwert λ mit $\mu = \lambda^2$ besitzt.

(6.4.18) **Satz** *(optimale Wahl von ω). G sei konsistent geordnet. Die Eigenwerte der Matrix A_{Ges} des zugehörigen Gesamtschrittverfahrens seien sämtlich reell, und ihr Spektralradius $\rho := \rho(A_{Ges}) < 1$. Es bezeichne*

$$\omega_{opt} := \frac{2}{1 + \sqrt{1 - \rho^2}} .$$

Wir behaupten:

(i) $$\rho(A(\omega)) = \begin{cases} \left(\frac{\rho\omega}{2} + \frac{1}{2} \sqrt{\rho^2\omega^2 - 4(\omega - 1)} \right)^2 , & \text{falls } 0 < \omega \leq \omega_{opt} , \\ \omega - 1 , & \text{falls } \omega_{opt} \leq \omega < 2 , \end{cases}$$

(ii) $$\rho(A(\omega_{opt})) = \inf \{\rho(A(\omega)) : 0 < \omega < 2\} ,$$

(iii) $$\rho_{opt} := \rho(A(\omega_{opt})) = \omega_{opt} - 1 = \frac{1 - \sqrt{1 - \rho^2}}{1 + \sqrt{1 - \rho^2}} .$$

Beweis. Es sei $0 < \omega < 2$, $\mu \in \mathbb{C}$, $\kappa := \sqrt{\mu}$ eine beliebige der beiden möglichen Wurzeln aus μ. Aus Hilfssatz (6.4.13) folgt, daß μ genau dann Eigenwert von $A(\omega)$ ist, wenn ein Eigenwert λ von A_{Ges} existiert mit

$$\kappa\lambda\omega = \kappa^2 + \omega - 1 ;$$

hierzu ist wiederum zu beachten, daß mit λ auch $-\lambda$ Eigenwert von A_{Ges} ist. Als Lösungen dieser Gleichung bezüglich κ erhält man

$$\text{(6.4.19)} \qquad \kappa_{1,2}(\lambda, \omega) = \frac{\lambda\omega}{2} \pm \frac{1}{2}\sqrt{\lambda^2\omega^2 - 4(\omega - 1)} .$$

Danach ist μ genau dann Eigenwert von $A(\omega)$, wenn es einen Eigenwert λ von A_{Ges} mit $\kappa = \kappa_1(\lambda, \omega)$ oder $\kappa = \kappa_2(\lambda, \omega)$ gibt. Es folgt

$$\rho(A(\omega)) = (\max\{|\kappa_i(\lambda, \omega)| : i = 1, 2;\ \lambda \text{ Eigenwert von } A_{Ges}\})^2 .$$

Wegen $|\kappa_1(-\lambda, \omega)| = |\kappa_2(\lambda, \omega)|$ genügt es, das Maximum nur über die nichtnegativen Eigenwerte λ von A_{Ges} zu bilden. Beachtet man weiter, daß für $\lambda^2\omega^2 - 4(\omega - 1) \geqq 0$

$$\text{(6.4.20)} \qquad |\kappa_1(\lambda, \omega)| = \frac{\lambda\omega}{2} + \frac{1}{2}\sqrt{\lambda^2\omega^2 - 4(\omega - 1)}$$

$$\geqq \frac{\lambda\omega}{2} - \frac{1}{2}\sqrt{\lambda^2\omega^2 - 4(\omega - 1)} \geqq 0 ,$$

mithin

$$\text{(6.4.21)} \qquad |\kappa_1(\lambda, \omega)| \geqq |\kappa_2(\lambda, \omega)|$$

und für $\lambda^2\omega^2 - 4(\omega - 1) \leqq 0$ – hier ist notwendig $\omega \geqq 1$ –

$$\text{(6.4.22)} \qquad \kappa_{1,2}(\lambda, \omega) = \frac{\lambda\omega}{2} \pm \frac{i}{2}\sqrt{4(\omega - 1) - \lambda^2\omega^2} ,$$

folglich

$$\text{(6.4.23)} \qquad |\kappa_1(\lambda, \omega)| = |\kappa_2(\lambda, \omega)| = \sqrt{\omega - 1}$$

zutrifft, so erkennt man, daß sogar

$$\text{(6.4.24)} \qquad \rho(A(\omega)) = (\max\{|\kappa_1(\lambda, \omega)| : \lambda \text{ Eigenwert von } A_{Ges},\ \lambda \geqq 0\})^2$$

gilt.

ω_{opt} $(\geqq 1)$ ist offenbar Nullstelle der quadratischen Gleichung

$$\rho^2\omega^2 - 4(\omega - 1) = 0 ;$$

ferner stellt man fest, daß

$$\text{(6.4.25)} \qquad \rho^2\omega^2 - 4(\omega - 1) \begin{cases} \leqq 0, & \text{falls } \omega_{opt} \leqq \omega < 2 , \\ > 0, & \text{falls } 0 < \omega < \omega_{opt} . \end{cases}$$

Hieraus folgt für $\omega_{opt} \leqq \omega < 2,\ 0 \leqq \lambda \leqq \rho$

$$\lambda^2\omega^2 - 4(\omega - 1) \leqq 0\,,$$

mithin nach (6.4.23) für diese λ, ω-Werte

$$|\kappa_1(\lambda, \omega)| = \sqrt{\omega - 1}\,.$$

Damit ist nach (6.4.24) für $\omega_{opt} \leqq \omega < 2$, wie behauptet, $\rho(A(\omega)) = \omega - 1$.

Im Fall $0 < \omega < \omega_{opt}$ bezeichne

$$\lambda_\omega := \min\{\lambda: 0 \leqq \lambda \leqq \rho: \lambda^2\omega^2 - 4(\omega - 1) \geqq 0\}\,.$$

In $[\lambda_\omega, \rho]$ ist nach (6.4.20) $|\kappa_1(\lambda, \omega)|$ bezüglich λ monoton wachsend, also

$$|\kappa_1(\lambda, \omega)| \leqq |\kappa_1(\rho, \omega)|\,.$$

Für $\lambda \in [0, \lambda_\omega[$ hat man nach (6.4.23)

$$|\kappa_1(\lambda, \omega)| = \sqrt{\omega - 1} < \frac{\rho\omega}{2} < |\kappa_1(\rho, \omega)|\,.$$

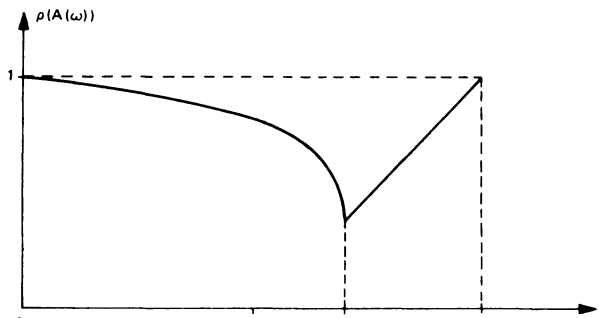
Insgesamt ist damit hier

$$|\kappa_1(\rho, \omega)| = \max\{|\kappa_1(\lambda, \omega)|: \lambda \text{ Eigenwert von } A_{Ges},\ \lambda \geqq 0\}$$

und somit die Behauptung (i) vollständig bewiesen.

Die Aussage (ii) ist klar, da nach (i) $\rho(A(\omega))$ in $]0, 2[$ stetig ist, in $]0, \omega_{opt}]$ streng monoton fällt und in $[\omega_{opt}, 2[$ streng monoton wächst. (iii) folgt aus (i) und (ii) durch Einsetzen.

Anmerkung. Der Aussage (i) entnimmt man – vgl. auch die untenstehende Zeichnung –, daß man ω lieber etwas größer als ω_{opt} wählen sollte als kleiner, wenn man ω_{opt} nicht genau bestimmen kann. Jedenfalls ist es unter den Voraussetzungen des Satzes (6.4.18) nicht sinnvoll, ein Verfahren der Unterrelaxation anzuwenden.



$\rho(A(\omega))$ im Fall $\rho = 0{,}9$.

6.5. Differenzenverfahren bei partiellen Differentialgleichungen

Die numerische Lösung von Randwertaufgaben bei partiellen Differentialgleichungen führt häufig auf lineare Gleichungssysteme, die nur iterativ zu lösen sind. Exemplarisch für derartige Probleme behandeln wir im Folgenden eine spezielle Randwertaufgabe im $\mathbb{R}^2$ über einem Quadrat.

Demgemäß gehen wir von

$$Q = [a,b] \times [a,b] \subset \mathbb{R}^2 ,$$
$$f : Q \to \mathbb{R}, \text{ stetig}$$

aus. Wir nennen – vgl. Abschnitt 6.6 – $\mathring{Q} :=]a,b[\times]a,b[$ das Innere, $\partial Q := Q \setminus \mathring{Q}$ den Rand von Q. Gesucht ist eine reellwertige Funktion $u(x,y)$, die in Q stetig, in $\mathring{Q}$ zweimal stetig differenzierbar ist und den Bedingungen

$$(6.5.1) \qquad \begin{cases} \Delta u := \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = f & \text{in } \mathring{Q}, \\ u = 0 & \text{auf } \partial Q \end{cases}$$

genügt.

Zur näherungsweisen Lösung dieser Randwertaufgabe ersetzt man den Differentialoperator Δ durch einen geeigneten Differenzenoperator Δ_h. Zur Motivation dieses Übergangs überlegen wir:

Für eine stetige Funktion z von $[a,b]$ in $\mathbb{R}$, die in $]a,b[$ zweimal stetig differenzierbar ist, gilt nach dem Satz von Taylor für $x_0 \in]a,b[$, $0 < h \leq \min\{b - x_0, x_0 - a\}$

$$(*) \qquad \begin{cases} z(x_0 + h) = z(x_0) + h \cdot z'(x_0) + \dfrac{h^2}{2} \cdot z''(\xi(h)) , \\ z(x_0 - h) = z(x_0) - h \cdot z'(x_0) + \dfrac{h^2}{2} \cdot z''(\eta(h)) \end{cases}$$

mit $\xi(h), \eta(h) \in]x_0 - h, x_0 + h[$. Aus der Stetigkeit von z'' folgt

$$z''(\xi(h)) + z''(\eta(h)) = 2 \cdot (z''(x_0) + \epsilon(h)) ,$$

wobei $\epsilon(h)$ eine Funktion ist, die mit $h \searrow 0$ gegen 0 strebt. Durch Addition der beiden Gleichungen in (*) gewinnen wir daher

$$z(x_0 + h) + z(x_0 - h) = 2\, z(x_0) + h^2 (z''(x_0) + \epsilon(h)) ,$$

also auch

$$(6.5.2) \qquad z''(x_0) = \lim_{h \searrow 0} \frac{1}{h^2} (z(x_0 + h) + z(x_0 - h) - 2z(x_0)) .$$

Hiervon ausgehend, definieren wir für $u : Q \to \mathbb{R}$, $(x,y) \in \mathring{Q}$, $h > 0$, genügend klein

$$(6.5.3) \qquad \Delta_h u(x,y) := \frac{1}{h^2}\{u(x+h,y)+u(x-h,y)+u(x,y+h)+u(x,y-h)-4u(x,y)\}$$

Für eine in Q stetige, in $\mathring{Q}$ zweimal stetig differenzierbare Funktion u folgt dann aus (6.5.2) unmittelbar

$$\lim_{h \searrow 0} \Delta_h u(x,y) = \Delta u(x,y) \qquad ((x,y) \in \mathring{Q}) .$$

Nach diesen Vorüberlegungen wählen wir zu einem $N \in \mathbb{N}$, $\geqq 2$

$$h = \frac{b-a}{N}$$

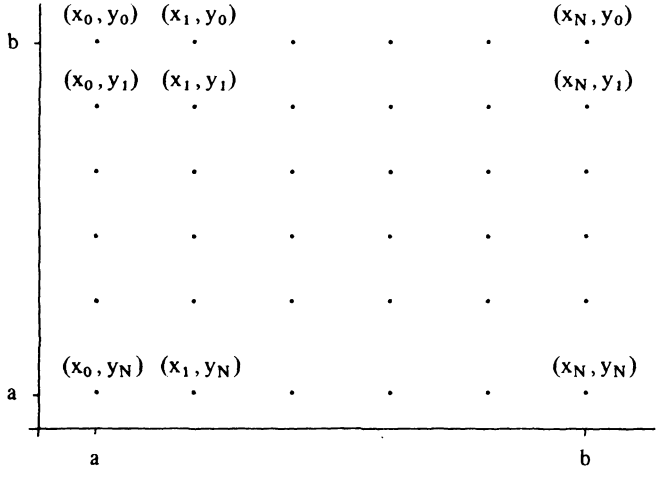
und definieren

$$(6.5.4) \qquad \begin{cases} x_\nu := a + \nu h & (\nu = 0, \ldots, N) , \\ y_\mu := b - \mu h & (\mu = 0, \ldots, N) . \end{cases}$$

Dann ist durch

$$(6.5.5) \qquad Q_h := \{(x_\nu, y_\mu) : \mu, \nu = 0, \ldots, N\}$$

eine endliche Teilmenge von Q erklärt; es handelt sich um ein quadratisches Punktgitter mit folgender Anordnung:



Hierzu bezeichnen wir

$$\mathring{Q}_h := \{(x_\nu, y_\mu) : \mu, \nu = 1, \ldots, N-1\} \quad (= \mathring{Q} \cap Q_h) ,$$

$$\partial Q_h := Q_h \setminus \mathring{Q}_h .$$

Als Näherung für die Lösung u von (6.5.1) bestimmen wir nun eine Abbildung $u_h : Q_h \to \mathbb{R}$ durch

$$(6.5.6)\qquad \begin{cases} \Delta_h u_h(x,y) = f(x,y) & ((x,y) \in \mathring{Q}_h)\,, \\ u_h(x,y) = 0 & ((x,y) \in \partial\mathring{Q}_h)\,. \end{cases}$$

Wenn f in $\mathring{Q}$ stetig differenzierbar ist, konvergieren die u_h für $h \searrow 0$ gegen die Lösung u von (6.5.1). Einen Beweis hierzu, mit dessen Hilfe auch die eindeutige Lösbarkeit von (6.5.1) gezeigt wird, findet man bei Walter [56]. Die Konvergenzordnung des Verfahrens (6.5.6) ist verhältnismäßig gering; die Übungsaufgabe 6.8 bringt ein Beispiel, in dem

$$\max_{(x,y)\in Q_h} |u_h(x,y) - u(x,y)| = O(h^2) \quad (h \to 0)$$

eintritt.

Im weiteren diskutieren wir die Lösung von (6.5.6) bei festem $h = \frac{b-a}{N}$. Zunächst ergibt sich nach (6.5.3) für $1 \leqq \mu, \nu \leqq N-1$

$$\Delta_h u_h(x_\nu, y_\mu) = \frac{1}{h^2}\{u_h(x_{\nu+1}, y_\mu) + u_h(x_{\nu-1}, y_\mu) + u_h(x_\nu, y_{\mu-1}) + u_h(x_\nu, y_{\mu+1}) - 4\,u_h(x_\nu, y_\mu)\}\,.$$

Wenn wir also

$$u_{\mu,\nu} := u_h(x_\nu, y_\mu) \qquad (\mu, \nu = 0, \ldots, N)\,,$$
$$f_{\mu,\nu} := f(x_\nu, y_\mu) \qquad (\mu, \nu = 1, \ldots, N-1)$$

bezeichnen, so erhält (6.5.6) die Gestalt

$$(6.5.7)\qquad \begin{cases} -u_{\mu-1,\nu} - u_{\mu,\nu-1} + 4\,u_{\mu,\nu} - u_{\mu+1,\nu} - u_{\mu,\nu+1} = -h^2 f_{\mu,\nu} \\ \qquad (\mu, \nu = 1, \ldots, N-1)\,, \\ u_{\mu,\nu} = 0 \;\; ((\mu,\nu) \in \{0,N\} \times \{0,1,\ldots,N\} \cup \{0,1,\ldots,N\} \times \{0,N\})\,. \end{cases}$$

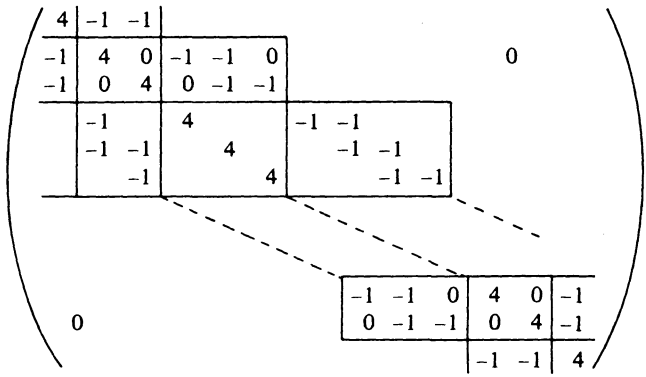
Demnach haben wir ein inhomogenes lineares System aus $(N-1)^2$ Gleichungen für die $(N-1)^2$ Unbekannten $u_{\mu,\nu}$ $(1 \leqq \mu, \nu \leqq N-1)$ zu lösen. Hierzu scheiden Eliminationsverfahren schon im Fall $N \approx 30$ aus, da sie zu viel Speicherplatz und Rechenzeit beanspruchen würden.

Wenn wir die Indexpaare (μ, ν) nach Schrägzeilen, also in der Reihenfolge

$$(6.5.8)\qquad \begin{cases} (1,1) & (1,2) & (1,3) \ldots\ldots & (1,N-1) \\ (2,1) & (2,2) & \ddots & \\ (3,1) & \ddots & & \\ \vdots & & & (N-2,N-1) \\ (N-1,1) & \ldots\ldots & (N-1,N-2) & (N-1,N-1) \end{cases}$$

anordnen und dann die $u_{\mu,\nu}$ in dieser Anordnung zu einem Vektor zusammenfassen sowie die Gleichungen (6.5.7) in entsprechender Weise untereinanderschreiben, so erhalten wir zu (6.5.7) die Koeffizientenmatrix

(6.5.9) $G =$



Offenbar erfüllt diese $((N-1)^2, (N-1)^2)$-Matrix das schwache Zeilensummenkriterium; daher ist nach (6.3.18) sowohl das Einzelschritt- als auch das Gesamtschrittverfahren konvergent.

Darüberhinaus sind für $0 < \omega < 2$ auch die Relaxationsverfahren (6.4.5) bezüglich G anwendbar. Zum Nachweis verifiziert man die Voraussetzungen des Satzes (6.4.18): Gemäß (6.2.20) gilt $\rho(A_{Ges}) < 1$. Wegen

(6.5.10) $A_{Ges} = I - \frac{1}{4} G$

ist mit G auch A_{Ges} symmetrisch; folglich sind die Eigenwerte von A_{Ges} sämtlich reell. Schließlich besitzt G die Eigenschaft A, ist also konsistent geordnet.

Im Folgenden bestimmen wir ω_{opt} und ρ_{opt} unter Anwendung des Satzes (6.4.18). Hierzu notieren wir zunächst den

(6.5.11) **Hilfssatz.** *Die Matrix des Gesamtschrittverfahrens, d.h. die Matrix* (6.5.10) *mit* G *aus* (6.5.9) *hat die Eigenwerte*

$$\lambda_{p,q} = \frac{1}{2}\left\{\cos\left(p\frac{\pi}{N}\right) + \cos\left(q\frac{\pi}{N}\right)\right\} \qquad (p, q = 1, \ldots, N-1)$$

und hierzu die paarweise orthogonalen Eigenvektoren

$$x_{p,q} = \left(\sin\left(p\frac{\mu\pi}{N}\right)\sin\left(q\frac{\nu\pi}{N}\right)\right)_{\mu,\nu=1}^{N-1} \qquad (\neq 0).$$

Den *Beweis* überlassen wir dem Leser als Übungsaufgabe 6.7.

Wegen der Orthogonalität der $x_{p,q}$ sind wir sicher, daß die angegebenen Eigenwerte $\lambda_{p,q}$ entsprechend ihrer Vielfachheit aufgezählt sind. Hieraus folgern wir

$$\rho(A_{Ges}) = \cos\frac{\pi}{N} \qquad (<1) \tag{6.5.12}$$

und weiter nach (6.4.17)

$$\rho(A_{Ein}) = \cos^2\frac{\pi}{N}\,; \tag{6.5.13}$$

schließlich liefert uns (6.4.18) für das Relaxationsverfahren die Werte

$$\begin{cases} \omega_{opt} = \dfrac{2}{1+\sin\frac{\pi}{N}}\,, \\[2ex] \rho_{opt} = \dfrac{1-\sin\frac{\pi}{N}}{1+\sin\frac{\pi}{N}}\,. \end{cases} \tag{6.5.14}$$

Zum Vergleich der Konvergenzgeschwindigkeiten bestimmen wir die Konvergenzverhältnisse $r(A) = -\ln\rho(A)$ – vgl. Abschnitt 6.2 – in 1. Näherung. Hierzu beachten wir, daß für hinreichend große N

$$\cos\frac{\pi}{N} \approx 1 - \frac{1}{2}\left(\frac{\pi}{N}\right)^2, \qquad \frac{1-\sin\frac{\pi}{N}}{1+\sin\frac{\pi}{N}} \approx 1 - 2\frac{\pi}{N}$$

und hinreichend kleine x

$$\log(1-x) \approx -x$$

gilt. Daher gewinnen wir die Näherungen

$$\begin{cases} r(A_{Ges}) \approx \dfrac{1}{2}\left(\dfrac{\pi}{N}\right)^2, \\[2ex] r(A_{Ein}) = 2\,r(A_{Ges}) \approx \left(\dfrac{\pi}{N}\right)^2, \\[2ex] r(A(\omega_{opt})) \approx 2\dfrac{\pi}{N} \approx \dfrac{4}{\pi}\cdot N \cdot r(A_{Ges})\,. \end{cases} \tag{6.5.15}$$

Das Relaxationsverfahren bei optimalem ω konvergiert also um etwa den Faktor N schneller als das Gesamtschrittverfahren.

Nach (6.3.4) lautet das Gesamtschrittverfahren

$$\begin{cases} u_{\mu,\nu}^{(m+1)} = \frac{1}{4}\,(u_{\mu,\nu-1}^{(m)} + u_{\mu-1,\nu}^{(m)} + u_{\mu,\nu+1}^{(m)} + u_{\mu+1,\nu}^{(m)} - h^2 f_{\mu,\nu})\,, \\ (\mu,\nu = 1,\ldots,N-1;\ m = 0,1,2,\ldots)\,. \end{cases} \tag{6.5.16}$$

und nach (6.4.6) das Relaxationsverfahren – bzw. für $\omega = 1$ das Einzelschrittverfahren –

$$(6.5.17)\quad \begin{cases} u_{\mu,\nu}^{(m+1)} = (1-\omega)u_{\mu,\nu}^{(m)} + \frac{\omega}{4}\left(u_{\mu,\nu-1}^{(m+1)} + u_{\mu-1,\nu}^{(m+1)} + u_{\mu,\nu+1}^{(m)} + u_{\mu+1,\nu}^{(m)} - h^2 f_{\mu,\nu}\right) \\ (\mu,\nu = 1,\ldots,N-1;\ m = 0,1,2,\ldots); \end{cases}$$

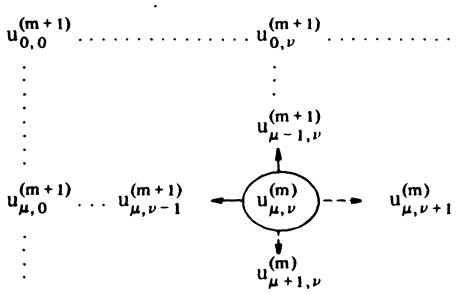
dabei sind für $m \in \mathbb{N}$ und

$$(\mu,\nu) \in \{0,N\} \times \{0,1,\ldots,N\} \cup \{0,1,\ldots,N\} \times \{0,N\}$$

die $u_{\mu,\nu}^{(m)} = 0$ zu setzen. Diese Größen werden zusätzlich eingeführt und wie angegeben festgelegt, um Ausnahmefälle bei der Notierung der Iterationsvorschriften (6.5.16), (6.5.17) zu vermeiden.

Gemäß der Definition der Matrix G sind die Indizes (μ,ν) bei den Rekursionen (6.5.16) und (6.5.17) wie in (6.5.8) notiert zu durchlaufen. Offenbar ist es möglich, diese Rekursionen auch in anderer Reihenfolge auszuwerten, sofern dabei sichergestellt ist, daß die jeweils rechts eingehenden Größen bereits berechnet sind. Dies ist, wie man leicht erkennt, z.B. dann der Fall, wenn man die (μ,ν) in (6.5.8) zeilenmäßig anordnet. Eine derartige Reihenfolge ist programmiertechnisch günstiger.

Entsprechend der geometrischen Anordnung der (x_ν, y_μ) in Q_h notiert man die $u_{\mu,\nu}^{(m+1)}$ in einem $(N+1, N+1)$-Matrixschema. Im Fall des Relaxationsverfahrens (6.5.17) läßt sich hierin die Rekursionsvorschrift folgendermaßen übersichtlich darstellen:



Man erkennt hieran, daß die Matrix der $u_{\mu,\nu}^{(m+1)}$ von links oben her aufzubauen ist, wie dies bei den vorgeschlagenen Berechnungsweisen auch geschieht.

Zur Fehlerabschätzung wollen wir die euklidische Norm im $\mathbb{C}^{(N-1)^2}$ heranziehen. Es ergibt sich für

$$\hat{u} := (u_{\mu,\nu})_{\mu,\nu=1}^{N-1},\ u_m := (u_{\mu,\nu}^{(m)})_{\mu,\nu=1}^{N-1} \quad (m \in \mathbb{N}),\ \geq 1)$$

im Fall des Gesamtschrittverfahrens nach (6.4.4) mit $\omega = 1$

$$|u_m - \hat{u}| \leqq |G^{-1}|\,|C_1 + C_2|\,|u_m - u_{m-1}| = 4\,|G^{-1}|\,|A_{Ges}|\,|u_m - u_{m-1}|$$
$$= 4 \cdot \rho(G^{-1}) \cdot \rho(A_{Ges})\,|u_m - u_{m-1}| .$$

Die letzte Gleichung folgt aus (3.3.12), da G^{-1} und A_{Ges} hermitesch sind. Wegen Hilfssatz (6.5.11) besitzt

$$G = 4\,(I - A_{Ges})$$

die Eigenwerte

$$\mu_{p,q} = 4 - 2\left(\cos\left(p\,\frac{\pi}{N}\right) + \cos\left(q\,\frac{\pi}{N}\right)\right) \qquad (p, q = 1, \ldots, N-1)$$

und hierzu die Eigenvektoren $x_{p,q}$. Es folgt

$$(6.5.18) \qquad \rho(G^{-1}) = \frac{1}{4\,(1 - \cos\frac{\pi}{N})} ,$$

also gewinnen wir für das Gesamtschrittverfahren die Abschätzung

$$(6.5.19) \qquad |u_m - \hat{u}| \leqq \frac{\cos\frac{\pi}{N}}{1 - \cos\frac{\pi}{N}}\,|u_m - u_{m-1}| \qquad (m \in \mathbb{N},\ \geqq 1) .$$

Im Fall des Einzelschritt- und Relaxationsverfahrens benutzen wir (6.4.8) mit $\omega = 1$ bzw. $\omega = \omega_{opt}$ $(\geqq 1)$. Wir erhalten

$$|u_m - \hat{u}| \leqq |G^{-1}|\left\{|\left(\frac{1}{\omega} - 1\right) D| + |C_2|\right\}|u_m - u_{m-1}|$$
$$= \rho(G^{-1})\left\{4\left(1 - \frac{1}{\omega}\right) + |C_2|\right\}|u_m - u_{m-1}| \qquad (m \in \mathbb{N},\ \geqq 1) .$$

Zur Abschätzung von $|C_2|$ beachten wir, daß durch C_2 die Abbildung

$$u = (u_{\mu,\nu})_{\mu,\nu=1}^{N-1} \mapsto (u_{\mu,\nu+1} + u_{\mu+1,\nu})_{\mu,\nu=1}^{N-1} =: v$$

(wobei $u_{\mu,N} = u_{N,\nu} = 0$ zu setzen ist) beschrieben wird. Eine elementare Abschätzung liefert

$$|v|^2 \leqq 4\,|u|^2 ,$$

mithin

$$(6.5.20) \qquad |C_2| \leqq 2 .$$

Wenn wir zusätzlich (6.5.18) benutzen, gewinnen wir für das Relaxationsverfahren mit $\omega = 1$ bzw. $\omega = \omega_{opt}$ die Ungleichungen

$$(6.5.21) \qquad |u_m - \hat{u}| \leqq \frac{1}{1 - \cos\frac{\pi}{N}}\left(\frac{3}{2} - \frac{1}{\omega}\right)|u_m - u_{m-1}| \qquad (m \in \mathbb{N},\ \geqq 1) .$$

Der Faktor $\left(\frac{3}{2} - \frac{1}{\omega}\right)$ ist für $\omega = \omega_{opt}$ größer als für $\omega = 1$, während das Verfahren mit $\omega = \omega_{opt}$ schneller konvergiert. Da sich also (6.5.21) im Fall $\omega = \omega_{opt}$ als zu grob erweist, wollen wir eine zweite Fehlerabschätzung angeben. Dazu schreiben wir (6.5.7) in der Form

$$G\hat{u} = c$$

und bestimmen für $m \in \mathbb{N}$ $r_m \in \mathbb{C}^{(N-1)^2}$ mit

$$G u_m = c + r_m .$$

Dann wird trivialerweise

$$(6.5.22) \qquad |u_m - \hat{u}| \leqq |G^{-1}|\,|r_m| = \frac{1}{4(1 - \cos\frac{\pi}{N})}\,|r_m| \quad (m \in \mathbb{N}) .$$

Selbstverständlich ist (6.5.22) auf jedes Iterationsverfahren anwendbar. Im Fall des Gesamtschrittverfahrens liefern (6.5.19) und (6.5.22) für große m asymptotisch gleiche Fehlerschranken. Vgl. hierzu Übungsaufgabe 6.9.

(6.5.23) **Zahlenbeispiel.** Wir betrachten in $Q = [-1, 1] \times [-1, 1]$ das folgende *Dirichletsche Randwertproblem*

$$(6.5.24) \qquad \begin{cases} \Delta v = 0 & \text{in } \mathring{Q} \\ v(x, y) = \varphi(x, y) := 1 + x + y^2 & ((x, y) \in \partial Q) . \end{cases}$$

Da im vorliegenden Fall die Funktion φ in ganz Q erklärt, dort stetig und in $\mathring{Q}$ zweimal stetig differenzierbar ist, erfüllt v die Aufgabe (6.5.24) genau dann, wenn

$$u := v - \varphi$$

das Randwertproblem

$$(6.5.25) \qquad \begin{cases} \Delta u = -\Delta\varphi \ (= -2) & \text{in } \mathring{Q} , \\ u = 0 & \text{auf } \partial Q \end{cases}$$

löst. Somit ist (6.5.24) auf ein Problem der Form (6.5.1) zurückgeführt. Wir wählen nun $h = 0{,}25$, also

$$N = 8$$

und lösen das Gleichungssystem (6.5.7) mittels der besprochenen Iterationsverfahren. Hierbei berechnen wir für jeden 5. Iterationsschritt die Fehlerschranken nach (6.5.19) bzw. (6.5.21) und nach (6.5.22); wir brechen ab, sobald wir auf diese Weise

$$(6.5.26) \qquad |u_m - \hat{u}| \leqq 2 \cdot 10^{-6}$$

festgestellt haben. In der folgenden Tabelle sind die Zahl der benötigten Iterationsschritte und die zugehörigen Fehlerschranken für die verschiedenen Iterationsverfahren einander gegenübergestellt.

	m	$\lvert u_m - \hat{u}\rvert \leq \ldots$ nach (6.5.19/21)	(6.5.22)
Gesamtschrittverfahren	180	$1{,}6778 \cdot 10^{-6}$	$1{,}6678 \cdot 10^{-6}$
Einzelschrittverfahren	90	$1{,}9437 \cdot 10^{-6}$	$2{,}0540 \cdot 10^{-6}$
Überrelaxation, $\omega = \omega_{opt}$	25	$3{,}0452 \cdot 10^{-6}$	$0{,}5424 \cdot 10^{-6}$

Ein Vergleich der m-Werte bestätigt, daß das Einzelschrittverfahren etwa zweimal, das Überrelaxationsverfahren etwa N-mal so schnell wie das Gesamtschrittverfahren konvergiert. Wie schon oben angekündigt, stimmen für das Gesamtschrittverfahren die nach (6.5.19) und (6.5.22) ermittelten Fehlerschranken nahezu überein.

Die Lösung $\hat{u}$ von (6.5.7) und damit u_h ist uns nunmehr auf mindestens 5 bis 6 Dezimalstellen hinter dem Komma bekannt. Als Näherungslösung von (6.5.24) gewinnen wir schließlich $v_h : Q_h \to \mathbb{R}$ mit

$$v_h(x, y) = u_h(x, y) + \varphi(x, y) \qquad ((x, y) \in Q_h) .$$

Wegen der großen Schrittweite h, die wir gewählt haben, um die Iterationsverfahren zu vergleichen, ist v_h allerdings eine ziemlich grobe Näherung von v. Nachfolgend sind die $v_h(x_\nu, y_\mu)$ $(\mu, \nu = 0, \ldots, N)$ in einer Matrix ausgedruckt.

1,0000	1,2500	1,5000	1,7500	2,0000	2,2500	2,5000	2,7500	3,0000
0,5625	0,9547	1,2845	1,5758	1,8387	2,0758	2,2845	2,4547	2,5625
0,2500	0,7220	1,1073	1,4301	1,7031	1,9301	2,1073	2,2220	2,2500
0,0625	0,5758	0,9926	1,3343	1,6135	1,8343	1,9926	2,0758	2,0625
0,0000	0,5262	0,9531	1,3010	1,5823	1,8010	1,9531	2,0262	2,0000
0,0625	0,5758	0,9926	1,3343	1,6135	1,8343	1,9926	2,0758	2,0625
0,2500	0,7220	1,1073	1,4301	1,7031	1,9301	2,1073	2,2220	2,2500
0,5625	0,9547	1,2845	1,5758	1,8387	2,0758	2,2845	2,4547	2,5625
1,0000	1,2500	1,5000	1,7500	2,0000	2,2500	2,5000	2,7500	3,0000

6.6. Differenzierbare Abbildungen, vereinfachtes Newton-Verfahren

Es seien R, S normierte Vektorräume über $\mathbb{K} = \mathbb{R}$ oder $\mathbb{C}$; – der Einfachheit halber unterlassen wir hier und i.a. auch im Folgenden die Unterscheidung verschiedener Normen durch Indizierung. D sei eine Teilmenge von R. Für $\rho > 0$ bezeichne – vgl. Abschnitt 6.1 –

$$K(x, \rho) := \{x' \in R : |x' - x| \,(\leqq)\, \rho\}$$

und hiermit

$$\mathring{D} := \{x \in D : \exists\, \rho > 0 \;\; K(x, \rho) \subset D\} ;$$

$\mathring{D}$ nennt man das *Innere von* D.

(6.6.1) **Definition.** f sei eine Abbildung von D in S; es sei $x_0 \in \mathring{D}$, $M \subset \mathring{D}$.
(i) f heißt differenzierbar in x_0 genau dann, wenn eine Abbildung $A \in L(R,S)$ und eine in x_0 stetige Abbildung ϵ von D in S mit $\epsilon(x_0) = 0$ existieren, so daß für alle $x \in D$

$$f(x) = f(x_0) + A(x - x_0) + |x - x_0|\,\epsilon(x)$$

gilt.
(ii) f heißt differenzierbar in M genau dann, wenn f in jedem $x_0 \in M$ differenzierbar ist.

Wir stellen fest: Ist f in x_0 differenzierbar, so sind die Abbildungen A, ϵ in (i) eindeutig bestimmt.

Zum *Beweis* nehmen wir an, es sei für $x \in D$ auch noch

$$f(x) = f(x_0) + B(x - x_0) + |x - x_0|\,\eta(x)\,,$$

wobei B und η die gleichen Eigenschaften wie A bzw. ϵ hätten. Durch Subtraktion ergibt sich dann für $x \in D, \neq x_0$

$$(A - B)\left(\frac{1}{|x-x_0|}(x-x_0)\right) = \eta(x) - \epsilon(x)\,,$$

Es sei nun $z \in R, \neq 0$. Setzt man

$$x_n = x_0 + \frac{1}{n} z \quad (n \in \mathbb{N}, \neq 0)\,,$$

so hat man

$$x_n \to x_0\ (n \to \infty),\ \frac{1}{|x_n - x_0|}(x_n - x_0) = \frac{1}{|z|} z \quad (n \in \mathbb{N}, \neq 0)\,.$$

Da x_0 im Inneren von D liegt, gehört somit x_n von einem n_0 ab zu D. Insgesamt ergibt sich infolgedessen

$$(A - B)\left(\frac{1}{|z|} z\right) = \eta(x_n) - \epsilon(x_n),\ \to 0 \quad (n \to \infty)\,,$$

mithin $(A - B)z = 0$, womit $A - B = 0$, d.h. $A = B$, also auch $\epsilon = \eta$ gezeigt ist.

Statt $A = A(x_0)$ schreibt man wie bei einer reellen Funktion einer reellen Veränderlichen $f'(x_0)$ oder $\frac{df}{dx}(x_0)$ und bezeichnet dies als *Ableitung der Abbildung* f *an der Stelle* x_0. Die Abbildung $f^{(1)}$ – oft auch als f' geschrieben –, definiert durch

$$D_{f^{(1)}} := \{x \in \mathring{D} : f \text{ differenzierbar in } x\}\,,$$
$$f^{(1)}(x) := f'(x) \quad (x \in D_{f^{(1)}})$$

nennt man die (erste) Ableitung von f.

(6.6.2) **Bemerkungen.**

(i) Ist f differenzierbar in x_0, so ist f dort stetig.

(ii) Ist mit $c \in S$ $f(x) = c$ für $x \in R$, d.h. f die konstante Abbildung c über R, so ist f in R differenzierbar und für $x_0 \in R$ $f'(x_0) = 0 \in L(R,S)$, also $f' = 0 \in L(R, L(R,S))$.

(iii) Ist mit $A \in L(R,S)$ $f(x) = Ax$ für $x \in R$, d.h. f die lineare beschränkte Abbildung A, so ist auch dieses f in R differenzierbar und für $x_0 \in R$ $f'(x_0) = A$, mithin $f^{(1)}$ über R die konstante Zuordnung $x \mapsto A$.

Beweis.

(i) Es sei $(x_n)_1^\infty$ eine Folge in D mit $x_n \to x_0$ $(n \to \infty)$. Wegen $f'(x_0) \in L(R,S)$ strebt für $n \to \infty$ $f'(x_0)(x_n - x_0) \to 0$ und daher

$$f(x_n) = f(x_0) + f'(x_0)(x_n - x_0) + |x_n - x_0|\,\epsilon(x_n) \to f(x_0)\,.$$

(ii) Es seien wieder x_0, $x \in R$ beliebig. Mit $A = 0 \in L(R,S)$ und $\epsilon(x) = 0 \in S$ für $x \in R$ ergibt sich trivialerweise

$$f(x) = f(x_0) + 0(x - x_0) + |x - x_0| \cdot 0 \quad (x \in R)\,.$$

(iii) Es seien wieder x_0, $x \in R$ beliebig. Hier ist mit dem vorgegebenen $A \in L(R,S)$ und $\epsilon(x) = 0 \in S$ für $x \in R$

$$f(x) = f(x_0) + A(x - x_0) + |x - x_0| \cdot 0 \quad (x \in R)\,,$$

womit auch diese Aussage bewiesen ist.

Eingehender betrachten wir den wichtigen Spezialfall $R = \mathbb{R}^m$, $S = \mathbb{R}^n$; diese Räume seien dabei z.B. jeweils mit der euklidischen Norm versehen. Es sei also f eine Abbildung von $D \subset \mathbb{R}^m$ in $\mathbb{R}^n$, d.h.

$$f(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{pmatrix}\,,$$

wobei die $f_i(x)$ reellwertige Funktionen von m reellen Veränderlichen $x = (\xi_j)_1^m$ sind.

Nach (3.3.2) gilt

$$L(\mathbb{R}^m, \mathbb{R}^n) = \mathrm{Hom}(\mathbb{R}^m, \mathbb{R}^n) \cong M(n \times m, \mathbb{R})\,;$$

daher kann im vorliegenden Fall in Definition (6.6.1), (i) die Aussage „$A \in L(R,S)$" durch „$A \in M(n \times m, \mathbb{R})$" ersetzt werden. Ist nun f in diesem Sinne in x_0 differen-

zierbar, so existieren in $x_0 = (\xi_i^0)_1^m$ sämtliche partiellen Ableitungen der Funktionen f_i, und es gilt gemäß obiger Isomorphie

$$(6.6.3) \qquad A = f'(x_0) = \begin{pmatrix} \frac{\partial f_1}{\partial \xi_1}(x_0) & \cdots & \frac{\partial f_1}{\partial \xi_m}(x_0) \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial \xi_1}(x_0) & \cdots & \frac{\partial f_n}{\partial \xi_m}(x_0) \end{pmatrix}$$

Zum *Beweis* kann man sich o.E. auf $n = 1$ und ohne wesentliche Einschränkung auf $m = 2$ beschränken. Dann hat man z.B. für $x = (\xi_1, \xi_2^0) \in D$ – im Argument von f schreibt man x stets als Zeile –

$$f(\xi_1, \xi_2^0) = f(\xi_1^0, \xi_2^0) + (\alpha_{1,1}, \alpha_{1,2}) \begin{pmatrix} \xi_1 - \xi_1^0 \\ 0 \end{pmatrix} + \sqrt{(\xi_1 - \xi_1^0)^2 + 0^2} \cdot \epsilon(\xi_1, \xi_2^0) ,$$

mithin

$$f(\xi_1, \xi_2^0) = f(\xi_1^0, \xi_2^0) + \alpha_{1,1}(\xi_1 - \xi_1^0) + |\xi_1 - \xi_1^0| \cdot \tilde{\epsilon}(\xi_1) ,$$

wobei die Funktion $\tilde{\epsilon}$ in ξ_1^0 stetig und $\tilde{\epsilon}(\xi_1^0) = 0$ ist. Demgemäß ist $f(x)$ in x_0 partiell nach ξ_1 differenzierbar und $\alpha_{1,1} = \frac{\partial f}{\partial \xi_1}(x_0)$.

Im Folgenden leiten wir die üblichen, mehr oder weniger elementaren Differentiationsregeln her.

(6.6.4) **Satz** *(Summenregel). Es seien* R, S *normierte Vektorräume über* $\mathbb{K}$, *ferner* $\alpha, \beta \in \mathbb{K}$ *sowie* f_1, f_2 *Abbildungen von* $D \subset R$ *in* S, *schließlich* $x_0 \in \mathring{D}$ *und* f_1, f_2 *differenzierbar in* x_0.

Unter diesen Voraussetzungen ist auch die Summe $\alpha f_1 + \beta f_2$, *definiert durch*

$$(\alpha f_1 + \beta f_2)(x) := \alpha f_1(x) + \beta f_2(x) \quad (x \in D) ,$$

in x_0 *differenzierbar, und es gilt*

$$(\alpha f_1 + \beta f_2)'(x_0) = \alpha f_1'(x_0) + \beta f_2'(x_0) .$$

Beweis. Auf Grund der Voraussetzungen gibt es in x_0 stetige Abbildungen ϵ_1, ϵ_2 von D in S mit $\epsilon_1(x_0) = \epsilon_2(x_0) = 0$ und

$$\left.\begin{aligned} f_1(x) &= f_1(x_0) + f_1'(x_0)(x - x_0) + |x - x_0|\epsilon_1(x) \\ f_2(x) &= f_2(x_0) + f_2'(x_0)(x - x_0) + |x - x_0|\epsilon_2(x) \end{aligned}\right\} \quad (x \in D) .$$

Dies führt für $x \in D$ unmittelbar zu

$$(\alpha f_1 + \beta f_2)(x) = (\alpha f_1 + \beta f_2)(x_0) + (\alpha f_1'(x_0) + \beta f_2'(x_0))(x - x_0) + |x - x_0|(\alpha\epsilon_1(x) + \beta\epsilon_2(x))$$

damit ist die Behauptung bereits bewiesen, da nach (3.2.10) $\alpha f_1'(x_0) + \beta f_2'(x_0)$ zu $L(R,S)$ gehört und $\alpha\epsilon_1 + \beta\epsilon_2$ wegen $\alpha\epsilon_1(x_0) + \beta\epsilon_2(x_0) = 0$ und

$$|\alpha\epsilon_1(x) + \beta\epsilon_2(x)| \leqq |\alpha|\,|\epsilon_1(x)| + |\beta|\,|\epsilon_2(x)|$$

in x_0 stetig ist.

Nahezu völlig analog beweist man – vgl. Übungsaufgabe 6.10 – den

(6.6.5) **Satz** *(Produktregel). Es seien* R, S_1, S_2, S_3 *normierte Vektorräume über* $\mathbb{K}$, *weiter* f, g *Abbildungen von* $D \subset R$ *in* $L(S_1, S_2)$ *bzw. in* $L(S_2, S_3)$, *ferner* $x_0 \in \mathring{D}$ *und* f, g *differenzierbar in* x_0.

Wir behaupten: dann ist auch das Produkt gf, *definiert durch*

$$(gf)(x) := g(x)\,f(x) \quad (x \in D),$$

in x_0 *differenzierbar, und es gilt für* $k \in R$

$$(gf)'(x_0)k = (g'(x_0)k)\,f(x_0) + g(x_0)\,(f'(x_0)k).$$

Dieser Satz beinhaltet natürlich u.a. eine Regel über die Differentiation des Produkts zweier (geeigneter) rechteckiger Matrizen, deren Elemente von einer oder mehreren reellen bzw. komplexen Variablen abhängen. Eine ähnliche Anwendung, nämlich die Differentiation der Inversen einer quadratischen Matrix, ermöglicht der folgende

(6.6.6) **Satz** *(Quotientenregel). Es sei* R *ein normierter Vektorraum,* S *ein Banachraum über* $\mathbb{K}$, *weiter sei* f *eine Abbildung von* $D \subset R$ *in* $L(S)$, $x_0 \in \mathring{D}$, $f(x_0) \in J(S)$ *und* f *differenzierbar in* x_0. *Schließlich bezeichne*

$$\begin{cases} D_g := \{x \in D : f(x) \in J(S)\}, \\ g(x) := f(x)^{-1} \quad (x \in D_g) \quad . \end{cases}$$

Wir behaupten: $x_0 \in \mathring{D}_g$, g *ist differenzierbar in* x_0, *und für* $k \in R$ *gilt*

$$g'(x_0)k = -f(x_0)^{-1}(f'(x_0)k)\,f(x_0)^{-1}.$$

Beweis. a) Wir zeigen zunächst: $x_0 \in \mathring{D}_g$, d.h.

$$\exists\,\rho > 0 \quad \forall\, x \in K(x_0, \rho) \quad f(x) \in J(S).$$

Hierzu wenden wir die Folgerung (3.2.15) bezüglich $A = f(x_0)$ und $B = f(x)$ an. Da f in x_0 differenzierbar und $f(x_0)$ bijektiv ist, gilt für $x \in D$

$$(6.6.7) \qquad f(x_0)^{-1}(f(x) - f(x_0)) = f(x_0)^{-1}\{f'(x_0)(x - x_0) + |x - x_0|\,\epsilon(x)\}.$$

Dabei ist ϵ in x_0 stetig und $\epsilon(x_0) = 0$; infolgedessen existiert ein $\rho_1 > 0$, so daß $K(x_0, \rho_1)$ in D liegt und für $x \in K(x_0, \rho_1)$

$$|\epsilon(x)| \leqq 1,$$

mithin nach (6.6.7)

$$|f(x_0)^{-1}(f(x)-f(x_0))| \leq |f(x_0)^{-1}|\,(|f'(x_0)|+1)\,|x-x_0|$$

abschätzbar ist. Wählt man nun

$$\rho = \min\left\{\rho_1, \frac{1}{|f(x_0)^{-1}|\,(|f'(x_0)|+1)}\right\},$$

so ist mit diesem ρ die Aussage a) klar.

b) Für $x \in K(x_0,\rho)$ gilt nach (3.2.15)

$$f(x)^{-1} = \sum_{\nu=0}^{\infty} \{-f(x_0)^{-1}(f(x)-f(x_0))\}^{\nu} f(x_0)^{-1},$$

folglich unter Berücksichtigung von (6.6.7)

$$\text{(6.6.8)} \quad \begin{cases} f(x)^{-1} = f(x_0)^{-1} - f(x_0)^{-1}(f'(x_0)(x-x_0))\,f(x_0)^{-1} \\ \quad - |x-x_0|\,f(x_0)^{-1}\epsilon(x)\,f(x_0)^{-1} + \sum_{\nu=2}^{\infty} \{-f(x_0)^{-1}(f(x)-f(x_0))\}^{\nu} f(x_0)^{-1}. \end{cases}$$

Unmittelbar einsichtig ist, daß durch

$$Ak := -f(x_0)^{-1}(f'(x_0)k)\,f(x_0)^{-1} \qquad (k \in R)$$

eine lineare beschränkte Abbildung von R in L(S) gegeben ist, d.h. man hat $A \in L(R, L(S))$. Definiert man weiter $\eta(x)$ durch

$$\eta(x) = \begin{cases} 0 \in L(S), & \text{falls } x = x_0, \\ \dfrac{1}{|x-x_0|}\{f(x)^{-1} - f(x_0)^{-1} - A(x-x_0)\}, & \text{falls } x \in D_g,\ \neq x_0, \end{cases}$$

so ergibt sich nach (6.6.7), (6.6.8) mit geeigneten $\gamma_1, \gamma_2 \geq 0$ für $x \in K(x_0,\rho)$

$$|\eta(x)| \leq \gamma_1 |\epsilon(x)| + \gamma_2 |x-x_0|$$

und daher $\eta(x) \to \eta(x_0) = 0$ für $x \to x_0$.

Insgesamt ist damit die Quotientenregel vollständig bewiesen.

(6.6.9) **Satz** *(Kettenregel). Es seien* R, S, T *normierte Vektorräume über* $\mathbb{K}$. f *sei eine Abbildung von* $D_1 \subset R$ *in* S *und* g *eine Abbildung von* $D_2 \subset S$ *in* T; *dabei sei* $f(D_1) \subset D_2$, $x_0 \in \mathring{D}_1$, $y_0 = f(x_0) \in \mathring{D}_2$ *sowie* f *differenzierbar in* x_0 *und* g *differenzierbar in* y_0.

Unter diesen Voraussetzungen ist auch die Abbildung $g \circ f$, *definiert durch*

$$(g \circ f)(x) := g(f(x)) \qquad (x \in D_1),$$

in x_0 *differenzierbar, und zwar mit der Ableitung*

$$(g \circ f)'(x_0) = g'(y_0)\,f'(x_0).$$

Beweis. Nach Voraussetzung gilt für $x \in D_1$

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + |x - x_0|\,\epsilon_1(x)$$

und für $y \in D_2$

$$g(y) = g(y_0) + g'(y_0)(y - y_0) + |y - y_0|\,\epsilon_2(y)\,;$$

dabei ist ϵ_1 in x_0 stetig mit $\epsilon_1(x_0) = 0$ und ϵ_2 in y_0 stetig mit $\epsilon_2(y_0) = 0$.

Hieraus ergibt sich insbesondere für $x \in D_1$, $y = f(x)$

$$(6.6.10)\quad \begin{cases} (g \circ f)(x) = (g \circ f)(x_0) + g'(y_0) f'(x_0)(x - x_0) \\ \qquad\qquad + \{|x - x_0|\, g'(y_0)\epsilon_1(x) + |y - y_0|\,\epsilon_2(y)\}\,. \end{cases}$$

Durch

$$Ak := g'(y_0) f'(x_0) k \quad (k \in R)$$

ist nach (3.1.20) eine Abbildung $A \in L(R, T)$ definiert. Setzt man mit diesem A für $x \in D_1$

$$(g \circ f)(x) =: (g \circ f)(x_0) + A(x - x_0) + |x - x_0|\,\epsilon(x)\,,$$

wobei zusätzlich $\epsilon(x_0) = 0$ $(\in T)$ vereinbart sei, so ist nach (6.6.10)

$$|\epsilon(x)| \leqq |g'(y_0)|\,|\epsilon_1(x)| + (|f'(x_0)| + |\epsilon_1(x)|)\,|(\epsilon_2 \circ f)(x)|\,,$$

womit wegen $\epsilon_1(x) \to \epsilon_1(x_0) = 0$, $f(x) \to f(x_0)$ für $x \to x_0$ und $\epsilon_2(y) \to \epsilon_2(y_0) = 0$ für $y \to y_0$ die Stetigkeit von ϵ in x_0 aufgezeigt ist.

Benutzen, jedoch nicht beweisen wollen wir hier den

(6.6.11) **Satz** *(Hahn-Banach). Es sei* R *ein normierter Vektorraum über* $\mathbb{K}$, $M \subset R$ *ein* $\mathbb{K}$*-Unterraum,* $f \in L(M, \mathbb{K})$, *d.h.* f *ein beschränktes lineares Funktional über* M.

Wir behaupten: es existiert eine normgleiche Fortsetzung g *von* f *auf* R, *d.h. es gibt ein* $g \in L(R, \mathbb{K})$ *mit* $g|_M = f$ *und* $|g| = |f|$.

Einen Beweis dieses Satzes findet man z.B. in [26], S. 33. Betrachtet man speziell zu einem $x_0 \in R$

$$M = \operatorname{span}(x_0)\,,$$
$$f(x) = \alpha\,|x_0|\,, \quad \text{falls} \quad x = \alpha x_0 \in M\,,$$

so erhält man als

(6.6.12) **Folgerung.** Es sei R ein normierter Vektorraum über $\mathbb{K}$ und $x_0 \in R$. Dann existiert ein $g \in L(R, \mathbb{K})$ mit $g(x_0) = |x_0|$ und $|g| \leqq 1$.

Übrigens ist diese Folgerung (und auch der Satz von Hahn-Banach) im Spezialfall des $\mathbb{K}^n$, versehen mit der gewichteten Maximumnorm oder einer ge-

wichteten p-Norm, unmittelbar einsichtig. So schließt man etwa im Falle der gewichteten Maximumnorm folgendermaßen: Ist $x_0 \in \mathbb{K}^n$, $= (\xi_j^0)_1^n$, $\neq 0$ und

$$\|x_0\|_w := \max_{j=1}^{n} \frac{|\xi_j^0|}{w_j} = \frac{|\xi_{j_0}^0|}{w_{j_0}}, \quad \xi_{j_0}^0 = |\xi_{j_0}^0| e^{i\varphi_0},$$

so setzt man mit

$$\eta_j = \begin{cases} 0 & (j \neq j_0), \\ \dfrac{e^{i\varphi_0}}{w_{j_0}} & (j = j_0) \end{cases} \qquad (j = 1, \ldots, n)$$

für $x \in \mathbb{K}^n$, $= (\xi_j)_1^n$

$$g(x) := \sum_{i=1}^{n} \bar{\eta}_j \xi_j = \frac{e^{-i\varphi_0}}{w_{j_0}} \xi_{j_0}.$$

Gemäß dieser Definition gilt dann natürlich $g(x_0) = |x_0|$ und $|g| = 1$.

Als Ersatz des Mittelwertsatzes der Differentialrechnung formulieren wir den

(6.6.13) **Satz.** *Es seien* R, S *normierte Vektorräume über* $\mathbb{K}$, g *sei eine Abbildung von* $D \subset R$ *in* S; *diese sei auf der Strecke*

$$\overline{x\, x_0} := \{x_0 + t(x - x_0) : t \in [0, 1]\}$$

differenzierbar – d.h. $\overline{x\, x_0} \in D_{g'}$ *– mit*

$$|g'(\xi)| \leqq M \quad (\xi \in \overline{x\, x_0}).$$

Unter diesen Voraussetzungen gilt die Abschätzung

$$|g(x) - g(x_0)| \leqq M |x - x_0|.$$

Beweis. Wir betrachten zunächst den Fall, daß R, S normierte Vektorräume über $\mathbb{R}$ sind. Nach (6.6.12) existiert ein $l \in L(S, \mathbb{R})$ mit

$$l(g(x) - g(x_0)) = |g(x) - g(x_0)|, \quad |l| \leqq 1.$$

Wir setzen

$$\varphi(t) := x_0 + t(x - x_0) \quad (t \in [0, 1])$$

und hiermit

$$F(t) := (l \circ g \circ \varphi)(t) \qquad (t \in [0, 1]).$$

Diese Funktion bildet $[0, 1]$ in $\mathbb{R}$ ab; nach der Kettenregel (6.6.9) ist sie dort differenzierbar mit der Ableitung

$$F'(t) = l\, g'(\varphi(t))\, \varphi'(t) = l\, g'(\varphi(t))\, (x - x_0).$$

Der Mittelwertsatz der Differentialrechnung sichert die Existenz eines $\tau \in]0, 1[$ mit

$$F(1) - F(0) = F'(\tau)(1 - 0).$$

Einsetzen führt wegen $\varphi(\tau) \in \overline{x\, x_0}$ zu

$$|g(x) - g(x_0)| = l\, g'(\varphi(\tau))(x - x_0) = |l\, g'(\varphi(\tau))(x - x_0)|$$
$$\leqq |l|\, |g'(\varphi(\tau))|\, |x - x_0| \leqq M\, |x - x_0|,$$

mithin zur Behauptung.

Sind R, S normierte Vektorräume über $\mathbb{C}$, so faßt man sie als normierte Vektorräume über $\mathbb{R}$ und $g'(\xi)$ als reell-lineare beschränkte Abbildung von R in S auf.

Nach diesen Vorbereitungen beweisen wir den folgenden allgemeinen

(6.6.14) **Satz** *(über das vereinfachte Newton-Verfahren). Es sei* R *ein Banachraum über* $\mathbb{K}$, $x_0 \in R$, $0 < \rho \leqq \infty$, $\sigma \geqq 0$, $\mu > 0$ *und* $0 \leqq p < 1$. *Es bezeichne*

$$K(x_0, \rho) = \{x \in R : |x - x_0| \,(\leqq)\, \rho\}.$$

g *sei eine differenzierbare Abbildung von* $K(x_0, \rho) \subset R$ *in* R *mit*

$$|g(x_0)| \leqq \sigma.$$

Weiter sei $A \in J(R)$ *und*

$$|A^{-1}| \leqq \frac{1}{\mu},\ |A - g'(x)| \leqq \mu p \quad (x \in K(x_0, \rho)).$$

Schließlich sei

$$\sigma \,(\leqq)\, \mu \rho (1 - p).$$

Wir behaupten:

(i) g *besitzt in* $K(x_0, \rho)$ *genau eine Nullstelle* $\hat{x}$.

(ii) *Die durch die Iterationsvorschrift*

$$x_{n+1} = x_n - A^{-1} g(x_n) \quad (n \in \mathbb{N})$$

definierte Folge $(x_n)_0^\infty$ *strebt für* $n \to \infty$ *gegen* $\hat{x}$.

(iii) *Es gelten die Abschätzungen*

$$(6.6.15) \quad |x_n - \hat{x}| \leqq \frac{p^n}{1 - p} |x_1 - x_0| \qquad (n \in \mathbb{N}),$$

$$(6.6.16) \quad |x_n - \hat{x}| \leqq \frac{p}{1 - p} |x_n - x_{n-1}| \qquad (n \in \mathbb{N}, \neq 0).$$

Zum *Beweis* wenden wir den Fixpunktsatz (6.1.7) an. Hierzu versehen wir R mit der durch die Norm erzeugten Metrik $d(x, y) = |x - y|$ und setzen

$$T(x) := x - A^{-1} g(x) \quad (x \in K(x_0, \rho)).$$

Nach der Kettenregel – man beachte die Bemerkung (6.6.2), (ii) – erhält man

$$T'(x) = \mathrm{id} - A^{-1} g'(x) = A^{-1}(A - g'(x)) ,$$

also

$$|T'(x)| \leqq |A^{-1}| \, |A - g'(x)| \leqq \frac{1}{\mu} \mu p = p ;$$

damit ist nach dem vorangehenden Satz $|T| \leqq p \, (< 1)$, d.h. T eine Kontraktion. Die Bedingung (6.1.9) ist wegen

$$d(x_0, T(x_0)) = |x_0 - T(x_0)| = |A^{-1} g(x_0)| \leqq \frac{1}{\mu} \sigma ,$$

mithin

$$\frac{1}{1 - |T|} d(x_0, T(x_0)) \leqq \frac{1}{1 - p} \frac{1}{\mu} \sigma \, (\leqq) \, \rho$$

erfüllt. Damit sind bezüglich T sämtliche Voraussetzungen des Fixpunktsatzes gegeben. Die Übertragung der dortigen Behauptungen auf die Abbildung g liefert unmittelbar unsere obigen Aussagen (i), (ii), (iii).

(6.6.17) **Beispiel.** Wir wenden das vereinfachte Newton-Verfahren an, um die in der Nähe von $x_0 = \begin{pmatrix} 1 \\ -0{,}5 \end{pmatrix}$ liegende Nullstelle von

$$g(\xi_1, \xi_2) = \begin{pmatrix} (1 + 0{,}05\,\xi_2)(1 - \xi_1) - 0{,}025\,\xi_2^2 \\ -0{,}5 - \xi_2 + 0{,}025\,(1 - \xi_1)^2 + 0{,}02\,\xi_2^2 \end{pmatrix}$$

– vgl. Beispiel (6.1.13) – zu berechnen. g ist in ganz $\mathbb{R}^2$ differenzierbar, als Ableitung an der Stelle $x = (\xi_1, \xi_2) \in \mathbb{R}^2$ erhalten wir

$$g'(x) = \begin{pmatrix} -(1 + 0{,}05\,\xi_2) & 0{,}05\,(1 - \xi_1 - \xi_2) \\ -0{,}05\,(1 - \xi_1) & -1 + 0{,}04\,\xi_2 \end{pmatrix} .$$

In den Voraussetzungen von Satz (6.6.14) ist A invertierbar und eine Näherung von $g'(x_0)$. Demgemäß diskutieren wir die Möglichkeiten

(i) $A = A_1 := g'(x_0) = \begin{pmatrix} -0{,}975 & 0{,}025 \\ 0 & -1{,}02 \end{pmatrix}$,

(ii) $A = A_2 := \begin{pmatrix} -0{,}975 & 0 \\ 0 & -1{,}02 \end{pmatrix}$,

(iii) $A = A_3 := -I$.

Wir legen die Norm

$$|x| := \max \{|\xi_1|, |\xi_2|\}$$

im $\mathbb{R}^2$ zugrunde und wählen in jedem Fall

$$\rho = 0{,}1$$

sowie

$$\sigma = |g(x_0)| = 0{,}00625 .$$

Im *Fall (i)* ergibt sich

$$\frac{1}{|A_1^{-1}|} = \frac{1{,}02 \cdot 0{,}975}{1{,}045} \geqq 0{,}95 =: \mu_1 ;$$

ferner erhalten wir für

$$(*) \quad x = x_0 + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \in K(x_0, \rho) ,$$

d.h. mit $|\epsilon_1|, |\epsilon_2| \leqq 0{,}1$ die Gleichung

$$A_1 - g'(x) = \begin{pmatrix} 0{,}05\,\epsilon_2 & 0{,}05\,\epsilon_2 + 0{,}05\,\epsilon_1 \\ -0{,}05\,\epsilon_1 & -0{,}04\,\epsilon_2 \end{pmatrix}$$

und folglich

$$|A_1 - g'(x)| \leqq 0{,}015 \quad (x \in K(x_0, \rho)) .$$

Mit

$$p_1 = 0{,}016$$

ist dann

$$|A_1 - g'(x)| \leqq \mu_1 p_1 \quad (= 0{,}0152)$$

sowie

$$\mu_1 \rho (1 - p_1) = 0{,}09348 > \sigma ,$$

insgesamt also die Voraussetzung von Satz (6.6.14) erfüllt.
Im *Fall (ii)* wird

$$\frac{1}{|A_2^{-1}|} = 0{,}975 =: \mu_2 ,$$

ferner gilt für x gemäß (*) die Gleichung

$$A_2 - g'(x) = \begin{pmatrix} 0{,}05\,\epsilon_2 & -0{,}025 + 0{,}05\,(\epsilon_1 + \epsilon_2) \\ -0{,}05\,\epsilon_1 & -0{,}04\,\epsilon_2 \end{pmatrix} ,$$

also

$$|A_2 - g'(x)| \leqq 0{,}04 \leqq \mu_2 p_2 \quad (x \in K(x_0, \rho)) ,$$

wenn man

$$p_2 = 0{,}042$$

wählt. Auch hier ist

$$\mu_2 \rho (1 - p_2) = 0{,}0934 > \sigma$$

und damit die Voraussetzung von Satz (6.6.14) erfüllt.

Im *Fall (iii)* wird $\mu_3 = 1$, ferner gilt für x gemäß (*)

$$A_3 - g'(x) = \begin{pmatrix} -0{,}025 + 0{,}05\,\epsilon_2 & -0{,}025 + 0{,}05\,(\epsilon_1 + \epsilon_2) \\ -0{,}05\,\epsilon_1 & 0{,}02 - 0{,}04\,\epsilon_2 \end{pmatrix}$$

und folglich

$$|A_3 - g'(x)| \leqq 0{,}065 =: p_3 = \mu_3 p_3 .$$

Schließlich wird

$$\mu_3 \rho(1 - p_3) = 0{,}0935 > \sigma .$$

Im Fall (iii) stimmt übrigens das vereinfachte Newton-Verfahren mit dem Iterationsverfahren aus Beispiel (6.1.13) überein.

Wir führen das vereinfachte Newton-Verfahren jeweils in REAL*8-Arithmetik durch und schätzen die Fehler wie im Beispiel (6.1.13) ab, wobei wir (6.6.15) bzw. (6.6.16) mit $n = 1$ und $p = p_i$ heranziehen. Hiermit erreichen wir in den Fällen (i), (ii), (iii) eine Abschätzung $|\tilde{x}_m - \hat{x}| \leqq 0{,}7 \cdot 10^{-15}$ nach $m = 5$ bzw. 7 bzw. 9 Iterationen.

6.7. Das Newton-Verfahren

Die Konvergenz des Newton-Verfahrens läßt sich – wie bereits festgestellt – ähnlich wie beim vereinfachten Newton-Verfahren durch Zurückführung auf den Fixpunktsatz (6.1.7) begründen. Dabei sind – vgl. Übungsaufgabe 6.11 – bezüglich der Abbildung

$$T(x) = x - g'(x)^{-1} g(x)$$

eine Reihe von Voraussetzungen zu machen. So ist vor allem sicherzustellen, daß T eine Kontraktion wird. Dazu fordert man global im betrachteten Bereich zweimalige Differenzierbarkeit von g (zur Definition siehe Abschnitt 6.8) sowie Abschätzungen für $g(x)$, $g'(x)^{-1}$ und $g''(x)$ derart, daß ähnlich wie im Beweis zu Satz (6.6.14)

$$|T'(x)| \leqq p < 1$$

wird.

Im Folgenden geben wir einen direkten, nicht auf den Fixpunktsatz zurückgreifenden Beweis der Konvergenz des Newton-Verfahrens; dieser Beweis geht im wesentlichen auf Kantorovich [35] zurück. Er ermöglicht es, die Voraussetzungen bezüglich $g(x)$ und $g'(x)^{-1}$ weitgehend lediglich lokal an der Stelle der Ausgangsnäherung $x = x_0$ zu fordern. Auch kommt man dabei mit einmaliger Differenzierbarkeit von g und einer (wenn auch globalen) Lipschitzbeschränktheit von g' aus. Dennoch erhält man Fehlerabschätzungen, die die gegenüber dem vereinfachten Newton-Verfahren schnellere, nämlich „quadratische" Konvergenz zum Ausdruck bringen.

Vorweg beweisen wir den folgenden

(6.7.1) **Hilfssatz.** *Es seien* R, S *normierte Vektorräume über* $\mathbb{K}$. g *sei eine Abbildung von* $D \subset R$ *in* S; *diese sei auf der Strecke* $\overline{x\,x_0}$ *stetig differenzierbar und genüge der Ungleichung*

$$(6.7.2) \qquad |g'(\xi) - g'(x_0)| \leqq \gamma\,|x - x_0| \qquad (\xi \in \overline{x\,x_0})\,.$$

Unter diesen Voraussetzungen ist

$$|g(x) - g(x_0) - g'(x_0)\,(x - x_0)| \leqq \frac{\gamma}{2}\,|x - x_0|^2$$

abschätzbar.

Beweis. Wir setzen o.E. – vgl. den Beweis zu Satz (6.6.13) – voraus, daß R, S normierte Vektorräume über $\mathbb{R}$ sind. Zur Abkürzung führen wir die Abbildung

$$G(\xi) := g(\xi) - g(x_0) - g'(x_0)\,(\xi - x_0)$$

ein. G ist ebenso wie g auf $\overline{x\,x_0}$ stetig differenzierbar mit

$$G'(\xi) = g'(\xi) - g'(x_0)\,.$$

Aufgrund der Folgerung (6.6.12) existiert ein $l \in L(S, \mathbb{R})$, so daß

$$l(G(x)) = |G(x)|\,, \quad |l| \leqq 1\,.$$

Mit $\varphi(t) = x_0 + t(x - x_0)$ setzen wir für $t \in [0, 1]$

$$F(t) := (l \circ G \circ \varphi)\,(t)\,.$$

Diese reellwertige Funktion F ist nach der Kettenregel (6.6.9) in $[0, 1]$ stetig differenzierbar und besitzt dort die Ableitung

$$F'(t) = l(G'(\varphi(t)))\,\varphi'(t) = l(g'(\varphi(t)) - g'(x_0))\,(x - x_0)\,.$$

Gemäß der Voraussetzung (6.7.2) schätzen wir danach

$$|F'(t)| \leqq |l|\,\gamma\,|\varphi(t) - x_0|\cdot|x - x_0| \leqq \gamma\,t\,|x - x_0|^2$$

ab. Insgesamt ergibt sich so, wenn wir noch $F(0) = l(G(x_0)) = 0$ berücksichtigen,

$$|G(x)| = F(1) = \int_0^1 F'(t)\,dt \leqq \gamma\,|x - x_0|^2 \cdot \int_0^1 t\,dt = \frac{\gamma}{2}\,|x - x_0|^2\,,$$

was zu zeigen war.

Wir notieren nun den

(6.7.3) **Satz** *(über das Newton-Verfahren). Es sei* R *Banach-Raum über* $\mathbb{K}$, $x_0 \in R$, $\rho > 0$ *und* $\kappa \geqq 0$. *Weiter bezeichne*

$$K := K(x_0, 2\rho) = \{x \in R : |x - x_0| \leqq 2\rho\}\,.$$

g sei eine differenzierbare Abbildung von $K \subset R$ *in* R; *dabei sei* $g'(x_0) \in J(R)$ *und*

$$\rho = |g'(x_0)^{-1} g(x_0)|$$

sowie für $x, y \in K$

$$|g'(x) - g'(y)| \leqq \kappa |x - y| .$$

Schließlich gelte

$$(6.7.4) \qquad h := |g'(x_0)^{-1}| \kappa \rho \leqq \frac{1}{2} .$$

Wir behaupten:

(i) *g besitzt in K genau eine Nullstelle* $\hat{x}$.

(ii) *Die Iterationsvorschrift*

$$(6.7.5) \qquad x_{n+1} = x_n - g'(x_n)^{-1} g(x_n) \quad (n \in \mathbb{N})$$

definiert eine Folge $(x_n)_0^\infty$, *die für* $n \to \infty$ *gegen* $\hat{x}$ *konvergiert.*

(iii) *Mit*

$$q := \frac{1}{2} \frac{h}{1-h} \quad \left(\leqq \frac{1}{2} \right)$$

gilt die a priori Abschätzung

$$(6.7.6) \qquad |x_n - \hat{x}| \leqq \frac{1}{1-q} \frac{\rho}{2^n} (2h)^{2^n - 1} \quad (n \in \mathbb{N})$$

sowie mit den in (6.7.24) *angegebenen* $0 \leqq q_n \leqq \frac{1}{2}$ *die a posteriori Abschätzung*

$$(6.7.7) \qquad |x_n - \hat{x}| \leqq \frac{1}{1-q_n} |g'(x_n)^{-1}| \frac{\kappa}{2} |x_n - x_{n-1}|^2 \quad (n \in \mathbb{N}, \neq 0) .$$

Beweis.

a) Wir zeigen zunächst, daß für alle $n \in \mathbb{N}$ x_n in K liegt und $g'(x_n)$ zu $J(R)$ gehört, womit die Existenz der Folge $(x_n)_0^\infty$ sichergestellt ist. Hierzu setzen wir

$$(6.7.8) \qquad \rho_n := |g'(x_n)^{-1} g(x_n)|, \quad h_n := |g'(x_n)^{-1}| \kappa \rho_n$$

sowie

$$(6.7.9) \qquad \begin{cases} K_n := \{x \in R : |x - x_n| \leqq 2\rho_n\} \quad (n \in \mathbb{N}) , \\ K_{-1} := K \end{cases}$$

und beweisen durch Induktion nach n simultan die Aussagen:

$$x_n \in K, \; g'(x_n) \in J(R), \; h_n \leqq \frac{1}{2}, \; K_n \subset K_{n-1} .$$

Für $n = 0$ sind diese Aussagen unmittelbar nach Voraussetzung bzw. per definitionem erfüllt. Daher steht nun der Schluß von n auf $n + 1$ an. Nach Induktionsannahme hat man $x_n \in K$, $g'(x_n) \in J(R)$; infolgedessen ist durch (6.7.5) x_{n+1} definiert und

$$|x_{n+1} - x_n| = |g'(x_n)^{-1} g(x_n)| = \rho_n \leqq 2\rho_n$$

abschätzbar. Danach liegt x_{n+1} in K_n, also erst recht in K. Dies führt nach Voraussetzung zu

$$|g'(x_{n+1}) - g'(x_n)| \leqq \kappa\, |x_{n+1} - x_n|$$

und weiter gemäß Induktionsvoraussetzung zu

$$|g'(x_n)^{-1}(g'(x_{n+1}) - g'(x_n))| \leqq |g'(x_n)^{-1}|\, \kappa\, \rho_n = h_n \quad (< 1)\,.$$

Anwendung der Folgerung (3.2.15) bezüglich $A = g'(x_n)$, $B = g'(x_{n+1})$ liefert die Aussage $g'(x_{n+1}) \in J(R)$ und darüber hinaus die Abschätzung

$$(6.7.10) \qquad |g'(x_{n+1})^{-1}| \leqq \sum_{\nu=0}^{\infty} h_n^{\nu}\, |g'(x_n)^{-1}| = \frac{1}{1-h_n}\, |g'(x_n)^{-1}|\,.$$

Nun zu $h_{n+1} \leqq \frac{1}{2}$. Auf Grund des vorangehenden Hilfssatzes ergibt sich

$$(6.7.11) \qquad |g(x_{n+1}) - g(x_n) - g'(x_n)\,(x_{n+1} - x_n)| \leqq \frac{\kappa}{2}\, |x_{n+1} - x_n|^2\,.$$

Die Definition von x_{n+1} impliziert trivialerweise die Gleichung

$$(6.7.12) \qquad g(x_n) + g'(x_n)\,(x_{n+1} - x_n) = 0\,,$$

womit dann aus (6.7.11) die Ungleichung

$$(6.7.13) \qquad |g(x_{n+1})| \leqq \frac{\kappa}{2}\, |x_{n+1} - x_n|^2 = \frac{\kappa}{2}\, \rho_n^2$$

folgt. Weiter erhält man gemäß (6.7.8) mit (6.7.10), (6.7.13) sowie der Induktionsvoraussetzung $h_n \leqq \frac{1}{2}$

$$(6.7.14) \qquad \rho_{n+1} \leqq |g'(x_{n+1})^{-1}|\, |g(x_{n+1})| \leqq \frac{1}{1-h_n}\, |g'(x_n)^{-1}|\, \frac{\kappa}{2}\, \rho_n^2$$

$$= \frac{1}{2}\, \frac{h_n}{1-h_n}\, \rho_n \leqq \frac{1}{2}\, \rho_n$$

und

$$(6.7.15) \qquad h_{n+1} \leqq \frac{1}{1-h_n}\, |g'(x_n)^{-1}|\, \kappa\, \frac{1}{2}\, \frac{h_n}{1-h_n} \cdot \rho_n$$

$$= \frac{1}{2} \left(\frac{h_n}{1-h_n} \right)^2 \leqq \frac{1}{2}\,.$$

Zu zeigen bleibt die Beziehung $K_{n+1} \subset K_n$. Ist $x \in K_{n+1}$, d.h. – vgl. (6.7.14) –

$$|x - x_{n+1}| \leqq 2 \cdot \rho_{n+1} \leqq \rho_n\,,$$

so gilt

$$|x - x_n| \leqq |x - x_{n+1}| + |x_{n+1} - x_n| \leqq \rho_n + \rho_n = 2 \cdot \rho_n\,,$$

mithin $x \in K_n$.

b) Als nächstes beweisen wir die Aussage:

$$\exists \hat{x} \in K \quad x_n \to \hat{x} \quad (n \to \infty) .$$

Hierzu stellen wir fest, daß sich aus (6.7.14) die Ungleichung

$$\rho_{n+\nu} \leqq \frac{1}{2^\nu} \rho_n \qquad (\nu \in \mathbb{N}) \tag{6.7.16}$$

induktiv herleiten läßt und damit

$$\begin{aligned} |x_{n+k} - x_n| &\leqq \sum_{\nu=0}^{k-1} |x_{n+\nu} - x_{n+\nu-1}| \\ &= \sum_{\nu=0}^{k-1} \rho_{n+\nu} \leqq \sum_{\nu=0}^{k-1} \frac{1}{2^\nu} \rho_n \leqq 2 \cdot \rho_n \end{aligned} \tag{6.7.17}$$

abschätzbar ist. Gemäß (6.7.16) strebt $\rho_n \to 0$ für $n \to \infty$; daher ist die Folge $(x_n)_0^\infty$ Cauchy-konvergent. Weil R vollständig und K abgeschlossen ist, existiert ein $\hat{x} \in K$ mit

$$\lim_{n \to \infty} x_n = \hat{x} .$$

c) Wir zeigen: $g(\hat{x}) = 0$. Da g in K differenzierbar, also sicherlich stetig ist, gilt trivialerweise

$$\lim_{n \to \infty} g(x_n) = g(\hat{x}) .$$

Weiter hat man nach Voraussetzung

$$|g'(x_n) - g'(\hat{x})| \leqq \kappa \, |x_n - \hat{x}| ,$$

mithin auch

$$\lim_{n \to \infty} g'(x_n) = g'(\hat{x}) .$$

Um einzusehen, daß $\hat{x}$ eine Nullstelle von g ist, bleibt in (6.7.12) der Grenzübergang $n \to \infty$ durchzuführen.

d) Nun zur Eindeutigkeit der Nullstelle $\hat{x}$: Hierzu betrachten wir ein beliebiges $\tilde{x} \in K$ mit $g(\tilde{x}) = 0$. Die Rekursionsvorschrift (6.7.5) liefert

$$\begin{aligned} |\tilde{x} - x_{n+1}| &= |\tilde{x} - x_n + g'(x_n)^{-1}(g(x_n) - g(\tilde{x}))| \\ &\leqq |g'(x_n)^{-1}| \, |g(\tilde{x}) - g(x_n) - g'(x_n)(\tilde{x} - x_n)| , \end{aligned}$$

was unter Benutzung des Hilfssatzes (6.7.1) zu

$$|\tilde{x} - x_{n+1}| \leqq |g'(x_n)^{-1}| \, \frac{\kappa}{2} \, |\tilde{x} - x_n|^2 \tag{6.7.18}$$

führt. Induktiv folgern wir die Abschätzung

$$(6.7.19) \qquad |\tilde{x} - x_n| \leqq \frac{\rho}{2^{n-1}} \quad (n \in \mathbb{N}) .$$

Für $n = 0$ ist diese Aussage wegen $\tilde{x} \in K$ klar. Zum Schluß von n auf $n+1$ schätzen wir nach (6.7.18) mit (6.7.10)

$$|\tilde{x} - x_{n+1}| \leqq 2^n |g'(x_0)^{-1}| \frac{\kappa}{2} \left(\frac{\rho}{2^{n-1}}\right)^2$$

$$= \frac{1}{2^{n-1}} |g'(x_0)^{-1}| \kappa \rho \cdot \rho = \frac{1}{2^{n-1}} h\rho \leqq \frac{\rho}{2^n}$$

ab. (6.7.19) impliziert die Eindeutigkeitsaussage

$$\tilde{x} = \lim_{n \to \infty} x_n = \hat{x} .$$

e) Schließlich kommen wir zu den Abschätzungen des Abbruchfehlers. Lassen wir in (6.7.17) $k \to \infty$ streben, so ergibt sich für $n \in \mathbb{N}$

$$(6.7.20) \qquad |\hat{x} - x_n| \leqq 2 \cdot \rho_n .$$

Hierin ist nach (6.7.8), (6.7.13)

$$(6.7.21) \qquad \rho_n \leqq |g'(x_n)^{-1}| \, |g(x_n)| \leqq |g'(x_n)^{-1}| \frac{\kappa}{2} |x_n - x_{n-1}|^2$$

abschätzbar, womit die a posteriori Abschätzung

$$(6.7.22) \qquad |\hat{x} - x_n| \leqq |g'(x_n)^{-1}| \, \kappa \, |x_n - x_{n-1}|^2$$

hergeleitet ist. Die im Satz aufgeführte a posteriori Abschätzung (6.7.7) ist etwa um den Faktor $\frac{1}{2}$ genauer. Zu ihrer Herleitung verschärfen wir die Abschätzungen (6.7.16) zu

$$(6.7.23) \qquad \rho_{n+\nu} \leqq q_n^{\nu} \rho_n \quad (\nu \in \mathbb{N}) ;$$

darin bezeichnet

$$(6.7.24) \qquad q_n := \frac{1}{2} \frac{h_n}{1 - h_n} .$$

Aus $h_0 = h$ folgt $q_0 = q$, ferner ist wegen $h_{n+1} \leqq h_n$ auch $q_{n+1} \leqq q_n \leqq \frac{1}{2}$. Die Ungleichung (6.7.23) ist für $\nu = 0$ klar; zum Schluß von ν auf $\nu + 1$ stützen wir uns auf (6.7.14) und auf die Monotonie der q_n:

$$\rho_{n+\nu+1} \leqq q_{n+\nu} \rho_{n+\nu} \leqq q_{n+\nu} q_n^{\nu} \rho_n \leqq q_n^{\nu+1} \rho_n .$$

In Analogie zu (6.7.17) erschließen wir

$$|x_{n+k} - x_n| \leqq \sum_{\nu=0}^{k-1} \rho_{n+\nu} \leqq \sum_{\nu=0}^{k-1} q_n^{\nu} \rho_n \leqq \frac{1}{1 - q_n} \rho_n ,$$

was mit $k \to \infty$ zu

$$(6.7.25) \qquad |\hat{x} - x_n| \leqq \frac{1}{1 - q_n} \rho_n$$

führt, womit unter Berücksichtigung von (6.7.21) auch (6.7.7) bewiesen ist.

Zum Schluß zur a priori Abschätzung (6.7.6): $h_n \leqq \frac{1}{2}$ impliziert

$$(6.7.26) \qquad \frac{1}{1 - h_n} \leqq 2$$

und weiter nach (6.7.15)

$$(2h_{n+1}) \leqq (2h_n)^2 .$$

Durch Induktion leitet man hieraus

$$(6.7.27) \qquad (2h_n) \leqq (2h_0)^{2^n} = (2h)^{2^n}$$

ab. Auf diese Ungleichung stützen wir uns zum Beweis der Abschätzung

$$(6.7.28) \qquad \rho_n \leqq \frac{\rho}{2^n} (2h)^{2^n - 1} \qquad (n \in \mathbb{N}) .$$

Für $n = 0$ ist diese per definitionem gültig. Zum Schluß von n auf $n + 1$ beachten wir, daß sich wegen $h_n \leqq \frac{1}{2}$ aus (6.7.14)

$$\rho_{n+1} \leqq h_n \rho_n$$

ergibt. Die Induktionsannahme zusammen mit (6.7.27) ermöglicht sodann die gewünschte Abschätzung

$$\rho_{n+1} \leqq \frac{1}{2} (2h)^{2^n} \frac{\rho}{2^n} (2h)^{2^n - 1} = \frac{\rho}{2^{n+1}} (2h)^{2^{n+1} - 1} .$$

Es bleibt (6.7.28) in (6.7.25) einzusetzen und $q_n \leqq q$ abzuschätzen.

Da sich (6.7.6) durch weitere Abschätzungen aus (6.7.7) ergibt, ist naturgemäß (6.7.7) die schärfere Schranke. Wie Beispiele zeigen, ist sie i.a. um einige Zehnerpotenzen schärfer. Vgl. hierzu auch Döring [10].

Die Fehlerschranken (6.7.6), (6.7.7) sind nützlich zur Beurteilung des Konvergenzverhaltens der durch (6.7.5) definierten Folge $(x_n)_0^\infty$; darüber hinaus sind sie sogar zur Fehleranalyse brauchbar, sofern die eingehenden numerischen Rechnungen mit hinreichend großer Stellenzahl ausgeführt werden, d.h. keine Rundungsfehler auftreten, die in der Größenordnung der Fehlerschranken liegen. Für eine Fehlerrechnung, die auch die Rundungsfehler berücksichtigt, ist die Abschätzung (6.7.7) weniger geeignet, da x_n nicht exakt bekannt ist und daher die Größen $g'(x_n)^{-1}$ und q_n nur schwer abzuschätzen sind. Hingegen läßt sich die Ungleichung (6.7.6) zu einer Fehleranalyse, wie sie ähnlich im Beispiel (6.1.13) durchgeführt wurde, heranziehen.

(6.7.29) **Beispiel.** Wir berechnen die in der Nähe von

$$\tilde{x}_0 = \begin{pmatrix} 1 \\ -0{,}5 \end{pmatrix}$$

liegende Nullstelle $\hat{x}$ der im Beispiel (6.6.17) angegebenen Abbildung g mit Hilfe des Newton-Verfahrens unter Benutzung der REAL*8-Arithmetik aus (1.1.16). Die so gewonnenen numerischen Werte nennen wir $\tilde{x}_n$ $(n = 0, 1, 2, \ldots)$. Wir führen bei jedem Iterationsschritt eine Fehlerabschätzung für $|\tilde{x}_n - \hat{x}|$ $(n \geqq 1)$ bezüglich der Maximumsnorm in folgender Weise durch.

Wir verwenden, ähnlich wie in Beispiel (6.1.13), die Größen

$$\bar{x}_n := \tilde{x}_{n-1} - g'(\tilde{x}_{n-1})^{-1} g(\tilde{x}_{n-1});$$

wir erhalten dann eine Schranke für $|\bar{x}_n - \hat{x}|$, indem wir (6.7.6) bezüglich $n = 1$ auf $\tilde{x}_{n-1}$ als x_0 und $\bar{x}_n$ als x_1 anwenden.

Wie aus den Ausführungen zu Beispiel (6.6.17) hervorgeht, haben wir mit $\kappa = 0{,}15$ für alle $x, y \in \mathbb{R}^2$

$$|g'(x) - g'(y)| \leqq \kappa |x - y| .$$

Eine Rundungsfehleranalyse liefert uns die Abschätzungen

$$|\mathrm{gl}(g'(\tilde{x}_{n-1})^{-1} g(\tilde{x}_{n-1})) - g'(\tilde{x}_{n-1})^{-1} g(\tilde{x}_{n-1})| \leqq 1{,}1\,\tau;$$

daher können wir

$$\rho = |g'(\tilde{x}_{n-1})^{-1} g(\tilde{x}_{n-1})|$$

durch die numerisch zu berechnende Größe

$$\tilde{\rho} := |\mathrm{gl}(g'(\tilde{x}_{n-1})^{-1} g(\tilde{x}_{n-1}))| + 1{,}1\,\tau$$

nach oben abschätzen. In ähnlicher Weise bestimmen wir ein $\tilde{\mu}$ mit

$$|g'(\tilde{x}_{n-1})^{-1}| \leqq \tilde{\mu} .$$

Dann gilt natürlich

$$h \leqq \tilde{h} := \tilde{\mu} \kappa \tilde{\rho} .$$

Wie wir nachprüfen, ist bei jedem Iterationsschritt

$$\tilde{h} \leqq \frac{1}{2},$$

daher ist die Anwendung von (6.7.6) gerechtfertigt. Es wird

$$q \leqq \tilde{q} := \frac{1}{2} \frac{\tilde{h}}{1 - \tilde{h}}$$

und hiermit

$$|\bar{x}_n - \hat{x}| \leqq \frac{1}{1 - \tilde{q}} \tilde{\rho} \tilde{h} .$$

Wie eine weitere Rundungsfehleranalyse ergibt, ist $|\tilde{x}_n - \bar{x}_n| \leqq 1{,}1\,\tau$ und daher schließlich

$$|\tilde{x}_n - \hat{x}| \leqq 1{,}1\,\tau + \frac{1}{1-\tilde{q}}\,\tilde{\rho}\,\tilde{h}\,.$$

In der folgenden Tabelle ist die so berechnete Fehlerabschätzung aufgeführt, wobei der (etwas zu große) Wert

$$\tau = \frac{1}{2}\cdot 10^{-15}$$

zugrunde gelegt ist.

n	$\tilde{x}_n$	$\lvert\tilde{x}_n - \hat{x}\rvert \leqq$
0	1,0000000000000000 − 0,5000000000000000	–
1	0,9937154348919055 − 0,4950980392156863	$0{,}623\cdot 10^{-5}$
2	0,9937164353405738 − 0,4950966000527695	$0{,}327\cdot 10^{-12}$
3	0,9937164353404487 − 0,4950966000527043	$0{,}550\cdot 10^{-15}$

Ein mehr qualitatives Kriterium über das Konvergenzverhalten des Newton-Verfahrens beinhaltet die folgende

(6.7.30) **Ergänzung.** Es sei R ein Banach-Raum über $\mathbb{K}$, g eine differenzierbare Abbildung von $D \subset R$ in R, ferner $\hat{x} \in D$, $g(\hat{x}) = 0$, $g'(\hat{x}) \in J(R)$ sowie mit einem $\kappa \geqq 0$

$$|g'(x) - g'(y)| \leqq \kappa\,|x - y| \qquad (x, y \in D)\,.$$

Dann existiert ein $r > 0$, so daß für alle $x_0 \in K(\hat{x}, r)$ bezüglich $\rho = |g'(x_0)^{-1} g(x_0)|$ sowie κ die Voraussetzungen des Konvergenzsatzes (6.7.3) erfüllt sind und die durch (6.7.5) definierte Folge $(x_n)_0^\infty$ gegen $\hat{x}$ konvergiert.

Beweis. Zur Abkürzung bezeichne

$$\omega := |g'(\hat{x})^{-1}| \quad (>0)\,;$$

ferner sei o.E. $\kappa > 0$ angenommen. Da g insbesondere in $\hat{x}$ differenzierbar ist, können wir ein $r' > 0$ mit $K(\hat{x}, r') \subset D$ fixieren; weiter setzen wir

$$(6.7.31) \qquad r := \frac{1}{4}\min\{r',\ (\omega\kappa)^{-1}\}\,.$$

Wir gehen nun von einem beliebigem $x_0 \in K(\hat{x}, r)$ aus. Nach (3.2.15) gehört auf Grund der Ungleichung

$$|g'(\hat{x})^{-1}|\ |g'(\hat{x}) - g'(x_0)| \leqq \omega\kappa\,|\hat{x} - x_0| \leqq \omega\kappa r < 1$$

$g'(x_0)$ zu J(R), und man hat

$$(6.7.32) \qquad |g'(x_0)^{-1}| \leqq \frac{\omega}{1 - \omega\kappa r}\,.$$

Als nächstes zeigen wir, daß $K(x_0, 2\rho) \subset K(\hat{x}, r')$, mithin g in $K(x_0, 2\rho)$ differenzierbar ist. Hierzu schätzen wir unter Anwendung des Hilfssatzes (6.7.1) bezüglich $\hat{x}, x_0$

$$|g(x_0) + g'(x_0)(\hat{x} - x_0)| \leqq \frac{\kappa}{2} |\hat{x} - x_0|^2$$

und damit

$$\begin{aligned} (6.7.33) \qquad \rho &= |(\hat{x} - x_0) - g'(x_0)^{-1} \{g(x_0) + g'(x_0)(\hat{x} - x_0)\}| \\ &\leqq |\hat{x} - x_0| + \frac{\omega\kappa}{2(1 - \omega\kappa r)} |\hat{x} - x_0|^2 \\ &\leqq \left(1 + \frac{\omega\kappa r}{2(1 - \omega\kappa r)}\right) r \leqq \frac{7}{6} r \end{aligned}$$

ab. Danach folgt aus $|x - x_0| \leqq 2\rho$ unmittelbar

$$|x - \hat{x}| \leqq |x - x_0| + |x_0 - \hat{x}| \leqq 2\rho + r \leqq \frac{10}{3} r \leqq r' .$$

Weiter erhält man auf Grund der Definition (6.7.31) und der Ungleichungen (6.7.32), (6.7.33) die Abschätzung

$$h = |g'(x_0)^{-1}| \kappa \rho \leqq \frac{7}{6} \frac{\omega\kappa r}{1 - \omega\kappa r} \leqq \frac{1}{2} ,$$

womit die Gültigkeit der Bedingung (6.7.4) aufgezeigt ist.

Insgesamt haben wir so erkannt, daß, wie behauptet, bezüglich x_0 sämtliche Voraussetzungen des Satzes (6.7.3) erfüllt sind. Schließlich bleibt zu überlegen, daß $\hat{x}$ in $K(x_0, 2\rho)$ liegt. Dazu schätzen wir – von der ersten Zeile in (6.7.33) ausgehend –

$$\begin{aligned} \rho &\geqq |\hat{x} - x_0| - \frac{\omega\kappa}{2(1 - \omega\kappa r)} |\hat{x} - x_0|^2 \geqq |\hat{x} - x_0| \left(1 - \frac{\omega\kappa r}{2(1 - \omega\kappa r)}\right) \\ &\geqq \frac{5}{6} |\hat{x} - x_0| \end{aligned}$$

ab. Ergänzend vermerken wir hierzu – vgl. Übungsaufgabe 6.14 –, daß $\hat{x}$ sogar im größeren Bereich $K(\hat{x}, 4r)$ die einzige Nullstelle von g ist.

Das Kriterium (6.7.30) besagt, daß das Newton-Verfahren im Falle einer „einfachen" Nullstelle, d.h. $g(\hat{x}) = 0$, $g'(\hat{x}) \in J(R)$ immer konvergiert, sofern nur der Startwert x_0 hinreichend nahe bei $\hat{x}$ liegt. Da in vielen Problemen eine geeignete Ausgangsnäherung sehr schwer zu bestimmen ist, verwendet man im Fall $R = \mathbb{R}^n$ gelegentlich ein modifiziertes Newton-Verfahren, das auch dann konvergiert, wenn der Startwert x_0 weiter, als in (6.7.30) angegeben, von $\hat{x}$ entfernt ist. Diese Modifikation lautet

$$x_{n+1} = x_n - \lambda_n g'(x_n)^{-1} g(x_n) \quad (n \in \mathbb{N}) ;$$

hierbei sind die $0 < \lambda_n \leq 1$ geeignet zu wählen, so daß unter anderem bezüglich der euklidischen Norm stets

$$|g(x_{n+1})| < |g(x_n)|$$

wird. Eine genaue Vorschrift zur Wahl der λ_n und einen Konvergenzbeweis für dieses Verfahren findet man bei Stoer [51], S. 236; der Grundgedanke ist in der Übungsaufgabe 6.16 festgehalten. Dieses modifizierte Newton-Verfahren ist jedoch auch nur lokal konvergent.

Abschließend beschäftigen wir uns noch mit der Methode der *nicht-linearen Behandlung einer Eigenwertaufgabe.* Dabei lassen wir etwas allgemeinere Eigenwertaufgaben als in Kapitel 5 zu. So gehen wir aus von Matrizen $A, B \in M(n \times n, \mathbb{C})$ und suchen $\lambda \in \mathbb{C}$ sowie $y \in \mathbb{C}^n, \neq 0$, so daß

$$Ay = \lambda By \tag{6.7.34}$$

gilt; jedes derartige λ nennen wir einen Eigenwert des Problems (6.7.34) und jedes zugehörige y eine Eigenlösung. In Kapitel 5 haben wir eingehend die speziellere Problemstellung mit $B = I$ studiert; der allgemeinere Fall läßt sich natürlich auf diesen Spezialfall reduzieren, falls B invertierbar ist.

Ist A invertierbar, B die Nullmatrix, so besitzt (6.7.34) keinen Eigenwert; im Fall $A = B = 0$ sind sämtliche $\lambda \in \mathbb{C}$ Eigenwerte. Derartige Ausnahmefälle sind, sofern B invertierbar ist, natürlich nicht möglich.

Ein $y \in \mathbb{C}^n, \neq 0$ nennen wir einen Hauptvektor, genauer einen Hauptvektor k-ter Stufe zu einem Eigenwert λ, wenn es im $\mathbb{C}^n$ k, jedoch nicht weniger als k Vektoren $y_0, y_1, \ldots, y_{k-1} = y$ mit

$$\begin{cases} (A - \lambda B)\, y_0 = 0\,, \\ (A - \lambda B)\, y_1 = B y_0 \neq 0\,, \\ \quad\vdots \\ (A - \lambda B)\, y \;\; = B y_{k-2} \neq 0 \end{cases}$$

gibt. Der Hauptraum zum Eigenwert λ, nämlich der von den Hauptvektoren aufgespannte Unterraum sei mit $H_\lambda(A,B)$ bezeichnet. Der Eigenraum

$$E_\lambda(A,B) := \{y \in \mathbb{C}^n : (A - \lambda B)y = 0\}$$

ist natürlich ein Unterraum des Hauptraumes $H_\lambda(A,B)$. Im Fall $B = I$ reduzieren sich offensichtlich die hier neu eingeführten Begriffe auf die in Abschnitt 5.1, insbesondere in Satz (5.1.7) aus der linearen Algebra zitierten Bezeichnungen.

Ziel ist es, das Eigenwertproblem (6.7.34) mittels des Newton-Verfahrens zu behandeln. Dazu faßt man $y \in \mathbb{C}^n$ sowie $\lambda \in \mathbb{C}$ zu

$$x = \begin{pmatrix} y \\ \lambda \end{pmatrix} \in \mathbb{C}^n \times \mathbb{C}$$

zusammen und definiert durch

$$(6.7.35) \qquad g(x) := \begin{pmatrix} Ay - \lambda By \\ s(x) - 1 \end{pmatrix},$$

worin s eine lineare Abbildung von $\mathbb{C}^n \times \mathbb{C}$ in $\mathbb{C}$ ist, eine nichtlineare Abbildung g von $\mathbb{C}^n \times \mathbb{C}$ in sich. Für

$$(6.7.36) \qquad \hat{x} = \begin{pmatrix} \hat{y} \\ \hat{\lambda} \end{pmatrix}$$

gilt dann offenbar

$$(6.7.37) \qquad g(\hat{x}) = 0 \iff \begin{cases} A\hat{y} = \hat{\lambda} B\hat{y}, \\ s(\hat{x}) = 1. \end{cases}$$

Dazu bemerken wir Folgendes: Würde man statt (6.7.35) den Ansatz

$$\tilde{g}(x) = \begin{pmatrix} Ay - \lambda By \\ 0 \end{pmatrix}$$

machen, so könnte man hierauf den Konvergenzsatz (6.7.3) nicht anwenden. Falls nämlich ein $\hat{x}$ der Gestalt (6.7.36) mit $\hat{y} \neq 0$ eine Nullstelle von $\tilde{g}$ wäre, so träfe dasselbe auch für beliebiges $\mu \in \mathbb{C}$ auf

$$\hat{x}_\mu = \begin{pmatrix} \hat{y} \\ \hat{\lambda} \end{pmatrix} + \mu \begin{pmatrix} \hat{y} \\ 0 \end{pmatrix} = \begin{pmatrix} (1+\mu)\hat{y} \\ \hat{\lambda} \end{pmatrix}$$

zu. Daher lägen dann entgegen der Eindeutigkeit von $\hat{x}$ gemäß Satz (6.7.3) in jeder Umgebung von $\hat{x}$ unendlich viele Nullstellen von $\tilde{g}$.

Derartiges wird durch die „Normierungsbedingung" $s(\hat{x}) = 1$ bzw. – meist ist s von λ unabhängig – $s(\hat{y}) = 1$ ausgeschlossen. Auch wird hierdurch verhindert, daß das Newton-Verfahren zwar konvergiert, jedoch den nicht brauchbaren Grenzwert $\hat{y} = 0$ liefert.

Zu beachten ist, daß s so gewählt werden muß, daß es einen Eigenvektor $\hat{y}$ zu $\hat{\lambda}$ mit $s(\hat{x}) \neq 0$ gibt. Häufig setzt man

$$s(x) = e_i^t x = \eta_i,$$

worin η_i die i-te Komponente von y bezeichnet. Dabei kann man i geeignet wählen, wenn man wenigstens einen Eigenvektor hinreichend genau kennt. Dies trifft z. B. im Fall $B = I$ zu, wenn man mit Hilfe der Potenzmethode unter den Voraussetzungen des Zusatzes (5.1.19) die Näherungen $\lambda_1^{(k)}$ und $\hat{y}_k$ für ein $k \geq k_0$ ermittelt hat. Man setzt dann $i = i_0$ und wählt als Ausgangsnäherung

$$x_0 = \begin{pmatrix} \hat{y}_k \\ \lambda_1^{(k)} \end{pmatrix}.$$

Zur Anwendung des Satzes (6.7.3) ist die Ableitung $F'(x)$ zu bestimmen. Mit

$$h = \begin{pmatrix} c \\ \epsilon \end{pmatrix} \in \mathbb{C}^n \times \mathbb{C}, \neq 0$$

berechnen wir hierzu

$$g(x+h) - g(x) = \begin{pmatrix} A(y+c) - (\lambda+\epsilon)\, B(y+c) - Ay + \lambda By \\ s(x+h) - s(x) \end{pmatrix}$$

$$= \begin{pmatrix} (A - \lambda B)\, c - \epsilon By \\ s(h) \end{pmatrix} + \begin{pmatrix} -\epsilon Bc \\ 0 \end{pmatrix}.$$

Wie man leicht nachprüft, ist

$$\Phi_x(h) := \begin{pmatrix} (A - \lambda B)\, c - \epsilon By \\ s(h) \end{pmatrix} \tag{6.7.38}$$

linear in h, d.h. $\Phi_x \in L(\mathbb{C}^n \times \mathbb{C})$. Weiter erhält man, falls die Norm im Produkt beispielsweise durch

$$|h| = \max \{|c|, |\epsilon|\} \tag{6.7.39}$$

definiert ist,

$$|\epsilon Bc| \leqq |\epsilon|\, |B|\, |c| \leqq |h|^2\, |B|,$$

woraus sofort folgt, daß die Abbildung

$$\epsilon(h) := \begin{cases} \dfrac{1}{|h|} \begin{pmatrix} -\epsilon Bc \\ 0 \end{pmatrix} & \text{für } h \neq 0, \\ 0 & \text{sonst} \end{cases}$$

mit $h \to 0$ gegen 0 konvergiert. Insgesamt haben wir so bewiesen, daß g in x differenzierbar ist und dort die Ableitung $g'(x)h = \Phi_x(h)$ besitzt.

Als nächstes zeigen wir, daß wir bezüglich (6.7.35) in Satz (6.7.3) $\kappa = 2|B|$ wählen können. Mit

$$x_1, x_2 \in \mathbb{C}^n, \quad x_1 = \begin{pmatrix} y_1 \\ \lambda_1 \end{pmatrix}, \quad x_2 = \begin{pmatrix} y_2 \\ \lambda_2 \end{pmatrix}$$

schätzen wir hierzu unter Benutzung der Darstellung (6.7.38)

$$\begin{aligned} |(g'(x_1) - g'(x_2))\, h| &= |(\lambda_1 - \lambda_2) Bc + \epsilon B(y_1 - y_2)| \\ &\leqq |\lambda_1 - \lambda_2|\, |B|\, |c| + |\epsilon|\, |B|\, |y_1 - y_2| \\ &\leqq 2|B|\, |x_1 - x_2|\, |h| \end{aligned}$$

ab; diese Ungleichung bedeutet, daß wie behauptet

$$|g'(x_1) - g'(x_2)| \leqq 2|B|\, |x_1 - x_2|$$

gilt.

Zum Abschluß der Überlegungen notieren wir im Hinblick auf die Ergänzung (6.7.30) die

(6.7.40) **Bemerkung.** Das lineare Funktional s in (6.7.35) sei von λ unabhängig, und zwar als s(y) vorgegeben. $\hat{x} \in \mathbb{C}^n \times \mathbb{C}$, aufgeteilt wie in (6.7.36), genüge der Gleichung

$$g(\hat{x}) = \begin{pmatrix} A\hat{y} - \hat{\lambda}B\hat{y} \\ s(\hat{y}) - 1 \end{pmatrix} = 0 .$$

Ferner setzen wir voraus, daß

$$B\hat{y} \neq 0 .$$

Wir behaupten: $g'(\hat{x})$ ist genau dann invertierbar, also $g'(\hat{x}) \in J(\mathbb{C}^n \times \mathbb{C})$, wenn $\hat{\lambda}$ ein Eigenwert der Ordnung 1 des Problems (6.7.34) ist, d.h. wenn

$$\dim H_{\hat{\lambda}}(A, B) = 1$$

gilt.

Beweis.

(i) Wir nehmen an, daß $g'(\hat{x})$ nicht invertierbar sei, also ein

$$h = \begin{pmatrix} c \\ \epsilon \end{pmatrix} \neq 0$$

existiere mit

$$(6.7.41) \qquad g'(\hat{x})h = \begin{pmatrix} (A - \hat{\lambda}B)c - \epsilon B\hat{y} \\ s(c) \end{pmatrix} = 0 .$$

Zunächst untersuchen wir den Fall $\epsilon = 0$; dies impliziert natürlich $c \neq 0$.
Aus (6.7.41) folgt unmittelbar

$$(6.7.42) \qquad (A - \hat{\lambda}B)c = 0, \quad s(c) = 0 .$$

Die erste Gleichung besagt, daß c eine Eigenlösung des Problems (6.7.34) ist, die zweite wegen $s(\hat{y}) = 1$, daß c und $\hat{y}$ linear unabhängig sind. Insgesamt hat man demgemäß

$$\dim E_{\hat{\lambda}}(A, B) \geq 2 .$$

Wir kommen zum Fall $\epsilon \neq 0$, in dem wir o.E. $\epsilon = 1$ annehmen können. Der Gleichung (6.7.41) entnehmen wir die Beziehungen

$$(A - \hat{\lambda}B)c = B\hat{y} \neq 0, \quad s(c) = 0 .$$

Gemäß Definition ist danach c ein Hauptvektor 2. Stufe zu $\hat{\lambda}$, der wieder wegen $s(\hat{y}) \neq 0$, $s(c) = 0$ von $\hat{y}$ unabhängig ist, was auch hier

$$(6.7.43) \qquad \dim H_{\hat{\lambda}}(A, B) \geq 2$$

bedeutet.

(ii) Umgekehrt gehen wir nun davon aus, daß (6.7.43) gelte, und diskutieren dabei zunächst den Fall, daß sogar

$$\dim E_{\hat{\lambda}}(A,B) \geq 2$$

sei. Es existiert also zu $\hat{\lambda}$ ein von $\hat{y}$ unabhängiger Eigenvektor d. Wir setzen

(6.7.44) $\quad c := d - s(d)\hat{y}$

und stellen die Gültigkeit der Gleichungen (6.7.42) sowie $c \neq 0$ fest. Bezüglich

$$h := \begin{pmatrix} c \\ 0 \end{pmatrix}$$

ist dann (6.7.41) erfüllt, d.h. es ist $g'(\hat{x})$ nicht invertierbar. Zu diesem Ergebnis gelangt man auch, wenn unter der Bedingung (6.7.43)

$$\dim E_{\hat{\lambda}}(A,B) = 1$$

gilt. Es existiert dann nämlich zu $\hat{\lambda}$ ein Hauptvektor 2. Stufe bezüglich $\hat{y}$, d.h. ein $d \in \mathbb{C}^n$ mit

$$(A - \hat{\lambda}B)d = B\hat{y} \neq 0 .$$

Von d ausgehend, definieren wir c wie in (6.7.44); neben $s(c) = 0$ folgert man

$$(A - \hat{\lambda}B)c = (A - \hat{\lambda}B)d - s(d)(A - \hat{\lambda}B)\hat{y} = B\hat{y}$$

und daher bezüglich

$$h := \begin{pmatrix} c \\ 1 \end{pmatrix}$$

auch hier die Beziehung (6.7.41).

Die Voraussetzung $B\hat{y} \neq 0$ ist notwendig, wie die Übungsaufgabe 6.17 zeigt.

(6.7.45) **Beispiel.** Wir verbessern den im Beispiel (5.1.20) berechneten Eigenwert und Eigenvektor durch Anwendung des Newton-Verfahrens bezüglich der Abbildung (6.7.35). Hierbei gehen wir von der mit der Potenzmethode ermittelten Näherung

$$x_0 = \begin{pmatrix} \hat{y}_8 \\ \lambda^{(8)} \end{pmatrix}$$

aus und wählen als lineares Funktional

$$s(x) = s(y) := \eta_3 \quad (y = (\eta_i)_1^3) .$$

Wie in der Übungsaufgabe 6.18 begründet, ist dann bei jedem Iterationsschritt ein lineares Gleichungssystem mit einer (3,3)-Koeffizientenmatrix zu lösen. Wir erhalten bei 16-stelliger Rechnung folgende Werte:

k	λ_k	y_k		
0	4,965368144189775	0,9385061519605530	0,7484251534530475	1,0
1	5,000003470790874	0,9374962335066954	0,7500137554250997	1,0
2	5,000000000009646	0,9374999999996115	0,7500000000119355	1,0
3	5,000000000000000	0,9375000000000000	0,7500000000000000	1,0

6.8. Höhere Ableitungen; Iterationsverfahren höherer Ordnung

Die Ausführungen hier über höhere Ableitungen stellen in erster Linie Ergänzungen zu den vorangehenden Abschnitten 6.6 und 6.7 dar. Von Bedeutung sind sie jedoch darüber hinaus für eine kompakte Darstellung der Theorie über die numerische Integration von Differentialgleichungen; vgl. hierzu Band 3.

Es seien wieder R, S normierte Vektorräume über $\mathbb{K} = \mathbb{R}$ oder $\mathbb{C}$, ferner f eine Abbildung von $D \subset R$ in S. Rekursiv definieren wir die höheren Ableitungen von f, indem wir setzen:

$$(6.8.1)\quad \begin{cases} D_{f^{(0)}} = D,\ f^{(0)} = f, \\ D_{f^{(n+1)}} = \{x \in D_{f^{(n)}} : f^{(n)} \text{ dfb. in } x\}, \\ f^{(n+1)}(x) = f^{(n)'}(x) \quad (x \in D_{f^{(n+1)}}) \\ (n = 0, 1, 2, \ldots). \end{cases}$$

$f^{(n)}$ bezeichnen wir als die n-te Ableitung von f; dabei schließen wir nicht aus, daß der Definitionsbereich $D_{f^{(n)}}$ gegebenenfalls die leere Menge ist.

Weiter nennen wir f in $x_0 \in R$ n-mal differenzierbar, wenn x_0 zu $D_{f^{(n)}}$ gehört. f heißt in x_0 n-mal stetig differenzierbar, wenn f in x_0 n-mal differenzierbar und $f^{(n)}$ in x_0 stetig ist. Ist $M \subset R$, so sagt man entsprechend: f ist in oder auf M n-mal differenzierbar, wenn

$$M \subset D_{f^{(n)}}$$

gilt, sowie, f ist in oder auf M n-mal stetig differenzierbar, wenn f in M n-mal differenzierbar und die n-te Ableitung dort stetig ist.

Die Definition (6.8.1) ergibt beispielsweise für $n = 1$

$$D_{f^{(2)}} = \{x \in D_{f'} : f' \text{ dfb. in } x\},\quad f^{(2)}(x) = (f')'(x) \quad (x \in D_{f^{(2)}}).$$

Hierin ist f' eine Abbildung von $D_{f'} \subset R$ in $L(R,S)$ und dementsprechend $f^{(2)}$ eine Abbildung von $D_{f^{(2)}} \subset R$ in $L(R, L(R,S))$. Um die Bildräume beliebiger n-ter Ableitungen angeben zu können, definieren wir rekursiv:

$$(6.8.2)\quad \begin{cases} L_0(R,S) = S, \\ L_{n+1}(R,S) = L(R, L_n(R,S)) \quad (n = 0, 1, 2, \ldots). \end{cases}$$

Danach ist $f^{(n)}$ offenbar eine Abbildung von $D_{f^{(n)}} \subset R$ in $L_n(R,S)$.

Eine Abbildung

$$A: \underbrace{R \times R \times \ldots \times R}_{\text{n-mal}} \to S$$

heißt n-fach-linear oder multilinear, falls sie in jeder Koordinate (bei festen übrigen Koordinaten) linear ist, d.h.

$$A(\ldots, \alpha x + \beta y, \ldots) = \alpha A(\ldots, x, \ldots) + \beta A(\ldots, y, \ldots)$$

gilt. Die Gesamtheit n-fach-linearer Abbildungen von R in S sei mit $M_n(R,S)$ bezeichnet; sie wird, wenn man Addition und Multiplikation mit Skalar in kanonischer Weise definiert, ein Vektorraum über $\mathbb{K}$.

Eine n-fach-lineare Abbildung nennt man beschränkt, falls ein $\gamma \geq 0$ existiert, so daß für beliebige $x_1, x_2, \ldots, x_n \in R$

$$|A(x_1, x_2, \ldots, x_n)| \leq \gamma |x_1| \cdot |x_2| \cdot \ldots \cdot |x_n|$$

abschätzbar ist. Ist A in diesem Sinne beschränkt, so ist $|A|$ – die Norm von A – das Infimum = Minimum aller derartigen Schranken. Schließlich sei $Mb_n(R,S)$ die Menge aller beschränkten n-fach-linearen Abbildungen von R nach S; dabei setzen wir zusätzlich $M_0(R,S) = Mb_0(R,S) = S$.

Wie man leicht nachprüft, ist $Mb_n(R,S)$ ein normierter Unterraum des Vektorraums $M_n(R,S)$. Darüber hinaus gilt

(6.8.3) **Hilfssatz.** *Die normierten Vektorräume* $L_n(R,S)$ *und* $Mb_n(R,S)$ *sind normisomorph.*

Beweis. Nachzuweisen ist die Existenz einer linearen bijektiven Abbildung Φ_n von $L_n(R,S)$ auf $Mb_n(R,S)$ mit $|\Phi_n(A)| = |A|$ für $A \in L_n(R,S)$. Hierzu gehen wir induktiv vor. Für $n = 0$ und $n = 1$ ist die Aussage trivialerweise wahr. Wir kommen daher sofort zum Schluß von $n-1$ auf n: Für $A \in L_n(R,S)$ sowie $x \in R$ gehört Ax zu $L_{n-1}(R,S)$; dementsprechend definieren wir für $A \in L_n(R,S)$

$$(6.8.4) \qquad \Phi_n(A)(x, h_2, \ldots, h_n) := \Phi_{n-1}(Ax)(h_2, \ldots, h_n) \quad (x, h_2, \ldots, h_n \in R).$$

$\Phi_n(A)$ ist linear in den Variablen $h_2, \ldots, h_n$, da $\Phi_{n-1}(Ax)$ nach Induktionsannahme zu $Mb_{n-1}(R,S)$ gehört; ferner ist $\Phi_n(A)$ linear in x, da sowohl A als auch Φ_{n-1} linear sind. Weiter gilt nach Induktionsvoraussetzung

$$|\Phi_{n-1}(Ax)|_{M_{n-1}} = |Ax|_{L_{n-1}},$$

womit man

$$\begin{aligned} |\Phi_n(A)(x, h_2, \ldots, h_n)| &\leq |\Phi_{n-1}(Ax)|_{M_{n-1}} \cdot |h_2| \cdot \ldots \cdot |h_n| \\ &= |Ax|_{L_{n-1}} \cdot |h_2| \cdot \ldots \cdot |h_n| \\ &\leq |A|_{L_n} \cdot |x| \cdot |h_2| \cdot \ldots \cdot |h_n|, \end{aligned}$$

also

$$(6.8.5) \qquad |\Phi_n(A)|_{M_n} \leq |A|_{L_n} \quad (< \infty)$$

abschätzen kann. Beachtet man noch, daß aus der Linearität von Φ_{n-1} gemäß (6.8.4) sofort die Linearität von Φ_n folgt, so hat man bereits erkannt, daß Φ_n eine lineare beschränkte Abbildung von $L_n(R,S)$ in $Mb_n(R,S)$ ist.

Als nächstes überzeugen wir uns davon, daß Φ_n surjektiv ist. Es sei hierzu ein $B \in Mb_n(R,S)$ vorgegeben. Für $x \in R$ gehört $B(x,\ldots)$ zu $Mb_{n-1}(R,S)$ und infolgedessen nach Induktionsvoraussetzung

$$A(x) := \Phi_{n-1}^{-1}(B(x,\ldots)) \tag{6.8.6}$$

zu $L_{n-1}(R,S)$ mit

$$|A(x)|_{L_{n-1}} = |B(x,\ldots)|_{M_{n-1}}. \tag{6.8.7}$$

Die Linearität von Φ_{n-1}^{-1} und von B bezüglich x impliziert, daß A linear von x abhängt, d.h. eine lineare Abbildung von R in $L_{n-1}(R,S)$ ist. Um die Beschränktheit von A nachzuweisen, beachten wir, daß sich aus

$$|B(x,h_2,\ldots,h_n)| \leqq |B|_{M_n} \cdot |x| \cdot |h_2| \cdot \ldots \cdot |h_n| \quad (x,h_2,\ldots,h_n \in R)$$

mit (6.8.7) die Abschätzung

$$|Ax|_{L_{n-1}} = |B(x,\ldots)|_{M_{n-1}} \leqq |B|_{M_n} \cdot |x|,$$

das heißt

$$|A|_{L_n} \leqq |B|_{M_n} \tag{6.8.8}$$

ergibt. Danach gilt $A \in L_n(R,S)$, ferner gemäß (6.8.4), (6.8.6) $B = \Phi_n(A)$, womit die Surjektivität von Φ_n bewiesen ist.

Schließlich entnehmen wir den Ungleichungen (6.8.5), (6.8.8) die Normisomorphie

$$|\Phi_n(A)|_{M_n} = |A|_{L_n},$$

wonach gleichzeitig auch die Injektivität von Φ_n klar ist.

Der Hilfssatz (6.8.3) besagt, daß man die Elemente der Räume $L_n(R,S)$ und $Mb_n(R,S)$ identifizieren kann. Insbesondere werden wir dementsprechend im Folgenden die n-te Ableitung $f^{(n)}(x)$ als ein Element von $Mb_n(R,S)$ auffassen.

Wie bei der ersten Ableitung diskutieren wir auch bei der n-ten Ableitung eingehender den Spezialfall $R = \mathbb{R}^k$, $S = \mathbb{R}^l$:

Hierzu stellen wir zunächst fest: Ist $\dim R < \infty$, so gilt

$$Mb_n(R,S) = M_n(R,S); \tag{6.8.9}$$

diese Beziehung begründet man unter Benutzung der Hilfssätze (3.3.2) und (6.8.3), man kann sie jedoch auch direkt in Analogie zum Hilfssatz (3.3.2) erschließen.

Weiter bezeichne

$$T_n(k \times l; \mathbb{R}) := \{(\alpha_{i;j_1,\ldots,j_n})_{\substack{i=1,2,\ldots,l \\ j_\nu=1,2,\ldots,k \ (\nu=1,\ldots,n)}}\}$$

den Vektorraum der $k \times l$-Tensoren n-ter Stufe über $\mathbb{R}$. Wie aus der linearen Algebra bekannt ist, gilt

$$M_n(\mathbb{R}^k, \mathbb{R}^l) \cong T_n(k \times l; \mathbb{R}); \tag{6.8.10}$$

diese Isomorphie ist mit

$$h_\nu = \begin{pmatrix} \xi_1^{(\nu)} \\ \vdots \\ \xi_k^{(\nu)} \end{pmatrix} \in \mathbb{R}^k \qquad (\nu = 1, \ldots, n)$$

durch

$$Ah_1 \ldots h_n = \left(\sum_{j_1, \ldots, j_n = 1}^{k} \alpha_{i; j_1, \ldots, j_n} \xi_{j_1}^{(1)} \cdots \xi_{j_n}^{(n)} \right)_{i=1}^{l} \tag{6.8.11}$$

gegeben, worin bei vorgegebenem $A \in M_n(\mathbb{R}^k, \mathbb{R}^l)$ die Koeffizienten $\alpha_{i; j_1, \ldots, j_n}$ mit Hilfe der Einheitsvektoren e_j durch

$$\alpha_{i; j_1, \ldots, j_n} = e_i^t A e_{j_1} \cdots e_{j_n}$$

bestimmt sind. So ergibt sich z.B. im Fall $n = 2$, $l = 1$

$$Ah_1 h_2 = \sum_{j_1, j_2 = 1}^{k} \alpha_{j_1, j_2} \xi_{j_1}^{(1)} \xi_{j_2}^{(2)},$$

d.h. den Bilinearformen $A \in M_2(\mathbb{R}^k, \mathbb{R})$ sind die reellen (k,k)-Matrizen zugeordnet.

Es sei nun f eine Abbildung von $D \subset \mathbb{R}^k$ in $\mathbb{R}$, die in $x_0 \in D$ n-mal differenzierbar sei. Nach (6.8.3), (6.8.9) sowie (6.8.10) hat man

$$L_n(\mathbb{R}^k, \mathbb{R}) \cong Mb_n(\mathbb{R}^k, \mathbb{R}) = M_n(\mathbb{R}^k, \mathbb{R}) \cong T_n(k \times 1; \mathbb{R})$$

und daher im Sinne der Darstellung (6.8.11)

$$f^{(n)}(x_0) h_1 \ldots h_n = \sum_{j_1, \ldots, j_n = 1}^{k} f^{(n)}(x_0) e_{j_1} \ldots e_{j_n} \xi_{j_1}^{(1)} \cdots \xi_{j_n}^{(n)}. \tag{6.8.12}$$

Durch Induktion nach n zeigen wir, daß f in x_0 nach sämtlichen Koordinaten bis zur Ordnung n partiell differenzierbar ist und dabei

$$f^{(n)}(x_0) e_{j_1} \ldots e_{j_n} = \frac{\partial}{\partial \xi_{j_1}} \left(\cdots \frac{\partial}{\partial \xi_{j_{n-1}}} \left(\frac{\partial f}{\partial \xi_{j_n}} \right) \cdots \right)(x_0) \tag{6.8.13}$$

gilt. Ist $n = 1$, d.h. f in x_0 differenzierbar, so erhält man auf Grund der Kettenregel (6.6.9)

$$f'(x_0) e_{j_1} = \frac{d}{dt} f(x_0 + t e_{j_1}) \Big|_{t=0} = \lim_{t \to 0} \frac{f(x_0 + t e_{j_1}) - f(x_0)}{t} = \frac{\partial f}{\partial \xi_{j_1}}(x_0).$$

Zum Schluß von n auf n + 1 gehen wir davon aus, daß f in x_0 (n + 1)-mal differenzierbar ist. Setzen wir dann

$$D_g := D_{f^{(n)}},\ g(x) := f^{(n)}(x) \quad (x \in D_g),$$

so ist g eine Abbildung von $D_g \subset \mathbb{R}^k$ in $L_n(\mathbb{R}^k, \mathbb{R})$ sowie

$$x_0 \in \mathring{D}_g,\ g \text{ differenzierbar in } x_0,\ g'(x_0) = f^{(n+1)}(x_0).$$

Weiter hat man – man beachte dabei die Konvergenz bezüglich der Norm in $L_n(\mathbb{R}^k, \mathbb{R})$ bzw. $Mb_n(\mathbb{R}^k, \mathbb{R})$ –

$$(g'(x_0)e_{j_1})e_{j_2}\dots e_{j_n} = \left\{\lim_{t\to 0}\frac{g(x_0+te_{j_1})-g(x_0)}{t}\right\}e_{j_1}\dots e_{j_{n+1}}$$

$$= \lim_{t\to 0}\frac{1}{t}\{g(x_0+te_{j_1})e_{j_2}\dots e_{j_{n+1}} - g(x_0)e_{j_2}\dots e_{j_{n+1}}\}$$

und infolgedessen gemäß der Definition von g

$$f^{(n+1)}(x_0)e_{j_1}\dots e_{j_{n+1}} = \lim_{t\to 0}\frac{1}{t}\{f^{(n)}(x_0+te_{j_1})e_{j_2}\dots e_{j_{n+1}} - f^{(n)}(x_0)e_{j_2}\dots e_{j_{n+1}}\}.$$

Unter Benutzung der Induktionsannahme führt dies unmittelbar zu

$$f^{(n+1)}(x_0)e_{j_1}\dots e_{j_{n+1}} =$$

$$= \lim_{t\to 0}\frac{1}{t}\left\{\frac{\partial}{\partial\xi_{j_2}}\left(\dots\left(\frac{\partial f}{\partial\xi_{j_{n+1}}}\right)\dots\right)(x_0+te_{j_1}) - \frac{\partial}{\partial\xi_{j_2}}\left(\dots\left(\frac{\partial f}{\partial\xi_{j_{n+1}}}\right)\dots\right)(x_0)\right\}$$

$$= \left(\frac{\partial}{\partial\xi_{j_1}}\left(\frac{\partial}{\partial\xi_{j_2}}\dots\left(\frac{\partial f}{\partial\xi_{j_{n+1}}}\right)\dots\right)\right)(x_0),$$

was zu beweisen war.

Eine Abbildung $A \in M_n(R,S)$ nennt man *symmetrisch*, falls für alle $h_1, \dots, h_n \in R$ und jede Permutation σ der Zahlen $\{1, 2, \dots, n\}$, d.h. für jedes $\sigma \in S_n$

$$Ah_1h_2\dots h_n = Ah_{\sigma(1)}h_{\sigma(2)}\dots h_{\sigma(n)}$$

gilt. Auskunft über die Symmetrie der n-ten Ableitung einer Abbildung gibt der

(6.8.14) **Satz** *(von Schwarz). Es seien* R,S *normierte Vektorräume über* $\mathbb{K}$, f *eine Abbildung von* $D \subset R$ *in* S, $x_0 \in D$ *und* f *in* x_0 n*-mal differenzierbar.*

Wir behaupten: $f^{(n)}(x_0)$, *als Element von* $Mb_n(R,S)$ *aufgefaßt, ist symmetrisch.*

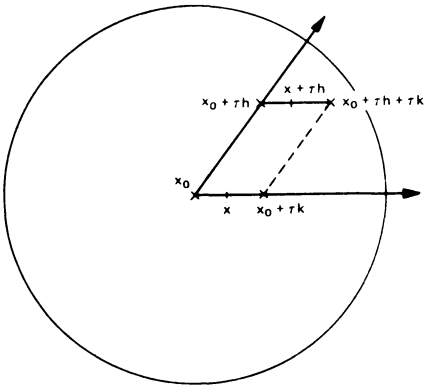
Beweis. Wir gehen induktiv vor. Für n = 1 ist nichts zu zeigen; der Kern des Beweises ist der Fall n = 2:

Hierzu wollen wir zunächst voraussetzen, daß R ein Vektorraum über $\mathbb{R}$ sowie S = $\mathbb{R}$ ist. Vorgegeben seien $h, k \in R$ und $\epsilon > 0$. Da f in x_0 zweimal

differenzierbar ist, können wir ein $\rho > 0$ wählen, so daß $K(x_0, \rho) \subset D_f$, und

(6.8.15) $\quad |f'(x) - f'(x_0) - f''(x_0)(x - x_0)| \leqq \epsilon \cdot |x - x_0| \quad (x \in K(x_0, \rho))$

gilt. Weiter bestimmen wir zu ρ ein $\tau > 0$ mit $\tau \cdot (|h| + |k|) < \rho$; offensichtlich gehören dann die Punkte $x_0 + \tau k$, $x_0 + \tau h$, $x_0 + \tau h + \tau k$ sowie für $x \in \overline{x_0, x_0 + \tau k}$ auch $x + \tau h$ zu $K(x_0, \rho)$.



Wir können daher

(6.8.16) $\quad g(x) = f(x + \tau h) - f(x) \quad (x \in \overline{x_0, x_0 + \tau k})$

sowie

(6.8.17) $\quad G(h, k)(\tau) = g(x_0 + \tau k) - g(x_0)$

setzen. Der Mittelwertsatz in $\mathbb{R}$, angewandt auf die reellwertige Funktion $h(t) := g(x_0 + tk)$ im Intervall $[0, \tau]$, liefert mit einem $0 < \xi < \tau$ unter Benutzung der Kettenregel (6.6.9)

$$g(x_0 + \tau k) - g(x_0) = g'(x_0 + \xi k)\tau k$$

und infolgedessen nach (6.8.16)

$$\begin{aligned} |G(h, k)(\tau) - \tau^2(f''(x_0)h)k| &= |g'(x_0 + \xi k)\tau k - \tau^2(f''(x_0)h)k| \\ &\leqq |g'(x_0 + \xi k) - f''(x_0)(\tau h)| \cdot \tau \cdot |k| \\ &\leqq \{|f'(x_0 + \tau h + \xi k) - f'(x_0) - f''(x_0)(\tau h + \xi k)| \\ &\quad + |f'(x_0 + \xi k) - f'(x_0) - f''(x_0)(\xi k)|\} \cdot \tau \cdot |k| . \end{aligned}$$

Hierauf wenden wir die Abschätzung (6.8.15) bezüglich $x = x_0 + \tau h + \xi k$ und

$x = x_0 + \xi k$ an, was zu

$$|G(h,k)(\tau) - \tau^2(f''(x_0)h)k| \leqq \epsilon\{|\tau h + \xi k| + |\xi k|\}\tau|k|$$
$$\leqq \epsilon(|h| + 2|k|)\cdot\tau^2|k|,$$

mithin zu

$$\overline{\lim_{\tau \searrow 0}} \left| \frac{G(h,k)(\tau)}{\tau^2} - (f''(x_0)h)k \right| \leqq \epsilon(|h| + 2|k|)|k|$$

führt. Da hierin $\epsilon > 0$ beliebig ist, verschwindet der Limes superior, und es existiert folglich

$$\lim_{\tau \searrow 0} \frac{G(h,k)(\tau)}{\tau^2} = (f''(x_0)h)k .$$

Zu beachten bleibt, daß h und k in $G(h,k)(\tau)$ vertauschbar sind.

Ist S ebenso wie R ein reeller Vektorraum, so betrachten wir mit $l \in L_1(S, \mathbb{R})$ die reellwertige Abbildung

$$\tilde{f}(x) = (l \circ f)(x) = l(f(x)) \quad (x \in D) .$$

Nach der Kettenregel (6.6.9) sowie der Bemerkung (6.6.2), (iii) ist $\tilde{f}$ in einer Kugel um x_0 differenzierbar und hat dort die Ableitung

$$\tilde{f}'(x) = l\,f'(x) .$$

Die Produktregel (6.6.5) und die Bemerkung (6.6.2), (ii) besagen, daß $\tilde{f}$ in x_0 zweimal differenzierbar ist und die zweite Ableitung dort mit $h \in R$ die Darstellung

$$\tilde{f}''(x_0)h = l(f''(x_0)h)$$

besitzt. Nimmt man ein $k \in R$ hinzu, so ergibt sich

$$(\tilde{f}''(x_0)h)k = l((f''(x_0)h)k) .$$

Wie wir bereits bewiesen haben, ist $\tilde{f}''(x_0)$ symmetrisch, d.h. h und k auf der linken Seite der letzten Gleichung, mithin auch auf ihrer rechten Seite sind vertauschbar. Daher ist für jedes lineare beschränkte Funktional l auf S

$$l\{(f''(x_0)h)k - (f''(x_0)k)h\} = 0 ,$$

mithin auf Grund der Folgerung (6.6.12) auch $f''(x_0)$ symmetrisch.

Sind R, S Vektorräume über $\mathbb{C}$, so fassen wir diese, wie wir es schon mehrfach gemacht haben, als Vektorräume über $\mathbb{R}$ auf.

Damit ist der Fall $n = 2$ vollständig bewiesen; wir kommen dementsprechend zum Schluß von n auf $n + 1$: Es sei also f in x_0 $(n + 1)$-mal differenzierbar; zu zeigen ist

$$f^{(n+1)}(x_0)\ldots h'\ldots h''\ldots = f^{(n+1)}(x_0)\ldots h''\ldots h'\ldots .$$

Wir betrachten zunächst den Fall, daß h' und h'' nicht an der ersten Stelle der Argumentliste stehen. Hier setzen wir

$$g(x) = f^{(n)}(x) \quad (x \in D_{f^{(n)}})$$

und erhalten wiederum mit (6.6.9)

$$\begin{aligned} f^{(n+1)}(x_0)h_1 \dots h' \dots h'' \dots &= (g'(x_0)h_1) \dots h' \dots h'' \dots \\ &= \left(\lim_{t \to 0} \frac{1}{t} \{g(x_0 + th_1) - g(x_0)\} \right) \dots h' \dots h'' \dots \\ &= \lim_{t \to 0} \frac{1}{t} \{g(x_0 + th_1) \dots h' \dots h'' \dots - g(x_0) \dots h' \dots h'' \dots\}, \end{aligned}$$

wobei zu berücksichtigen ist, daß es sich um Konvergenz in $L_n(R,S)$ bzw. $Mb_n(R,S)$ handelt. Die Induktionsannahme ergibt dann

$$\begin{aligned} f^{(n+1)}(x_0)h_1 \dots h' \dots h'' \dots &= \lim_{t \to 0} \frac{1}{t} \{g(x_0 + th_1) \dots h'' \dots h' \dots - g(x_0) \dots h'' \dots h' \dots\} \\ &= (g'(x_0)h_1) \dots h'' \dots h' \dots \\ &= f^{(n+1)}(x_0)h_1 \dots h'' \dots h' \dots . \end{aligned}$$

Es bleibt die Situation $h' = h_1$, $h'' = h_2$ zu untersuchen. In diesem Fall betrachten wir

$$g(x) = f^{(n-1)}(x) \quad (x \in D_{f^{(n-1)}})$$

und erschließen mit Hilfe der bereits bewiesenen Aussage für $n = 2$

$$\begin{aligned} f^{(n+1)}(x_0)h'h'' \dots &= (g''(x_0)h'h'') \dots \\ &= (g''(x_0)h''h') \dots = f^{(n+1)}(x_0)h''h' \dots . \end{aligned}$$

Die Bedeutung des Satzes (6.8.14) erläutern wir im Spezialfall $R = \mathbb{R}^k$, $S = \mathbb{R}$ näher: Ist eine Abbildung f in $x_0 \in D \subset \mathbb{R}^k$ n-mal differenzierbar, so ist $f^{(n)}(x_0)$ insbesondere im Bezug auf die Einheitsvektoren e_{j_ν} $(\nu = 1, \dots, n)$ symmetrisch, was nach (6.8.13) besagt, daß die partiellen Ableitungen bis zur Ordnung n in beliebiger Weise vertauschbar sind.

Unser nächstes Ziel ist der Satz von Taylor für Abbildungen in normierten Vektorräumen. Hierzu ist die Einführung und Diskussion von Polynomen in derartigen Räumen sinnvoll. Wir formulieren dementsprechend den

(6.8.18) **Hilfssatz.** *Es seien* R,S *normierte Vektorräume über* $\mathbb{K}$. *Für* $\nu = 0, 1, \dots, n$ *habe man Abbildungen* $A_\nu \in L_\nu(R,S)$, *die – im Sinne von* (6.8.3) *als Elemente von* $Mb_n(R,S)$ *aufgefaßt – symmetrisch seien; dabei gelte speziell* $A_n \neq 0$.

Wir setzen für $x \in R$, $0 \leqq \mu \leqq \nu$

$$A_\nu x^\mu := A_\nu x \dots x ,$$

definieren hiermit als „Polynom n-ten Grades"

$$P(x) := \sum_{\nu=0}^{n} A_\nu x^\nu$$

und behaupten: P *ist als Abbildung von* R *in* S *beliebig oft differenzierbar und besitzt für* $\kappa \in \mathbb{N}$ *die Ableitungen*

$$P^{(\kappa)}(x) = \sum_{\nu=\kappa}^{n} \nu(\nu-1)\dots(\nu-\kappa+1) A_\nu x^{\nu-\kappa},$$

wobei für $\kappa > n$ *als Wert der „leeren Summe" die* $0 \in L_\kappa(R,S)$ *anzusehen ist.*

Beweis. Zu symmetrischem $A \in L_\nu(R,S)$ betrachten wir für $0 \leqq \mu \leqq \nu$

$$h_\mu(x) := A x^\mu \quad (x \in R).$$

Wir zeigen: h_μ als Abbildung von R in $L_{\nu-\mu}(R,S)$ ist differenzierbar und hat dort die Ableitung

$$(6.8.19) \qquad h'(x) = \begin{cases} 0 \in L_{\nu+1}(R,S) & \text{für } \mu = 0, \\ \mu A x^{\mu-1} & \text{für } 1 \leqq \mu \leqq \nu. \end{cases}$$

Den *Beweis* hierzu führen wir durch Induktion nach μ. Nach (6.6.2), (ii) ist die konstante Abbildung $h_0(x) = A$ in R differenzierbar mit einer Ableitung der Form (6.8.19). Zum Schluß von μ auf $\mu + 1$ ($\leqq \nu$) stützen wir uns auf die in der Übungsaufgabe 6.10 angegebene Folgerung aus dem Produktsatz. Dazu identifizieren wir $\tilde{f} = id_R$, $g = h$. $\tilde{f}$ ist eine lineare beschränkte Abbildung von R in sich; gemäß (6.6.2), (iii) ist sie in R differenzierbar mit $\tilde{f}'(x) = id_R$ ($x \in R$). g ist eine Abbildung von R in $L(R, L_{\nu-\mu-1}(R,S))$; nach Induktionsannahme ist sie in R differenzierbar, wobei ihre Ableitung durch (6.8.19) gegeben ist. Nach Aufgabe 6.10, (ii) ist dann das Produkt

$$h_{\mu+1}(x) = h_\mu(x)\,\tilde{f}(x)$$

in R differenzierbar, und es gilt mit $k \in R$

$$(6.8.20) \qquad h'_{\mu+1}(x)k = (h'_\mu(x)k)\,x + h_\mu(x)k.$$

Ist $\mu \geqq 1$, so ergibt sich durch Einsetzen von (6.8.19)

$$h'_{\mu+1}(x)k = (\mu A x^{\mu-1}k)\,x + A x^\mu k,$$

folglich auf Grund der Symmetrie der Abbildung A

$$h'_{\mu+1}(x)k = (\mu+1) A x^\mu k,$$

also, da hierin k beliebig ist, die Induktionsbehauptung. Diese folgert man im Fall $\mu = 0$ ebenfalls aus (6.8.20), indem man die Beziehungen $h_0(x) = A$, $h'_0(x) = 0$ beachtet.

Wir beweisen nun in Verallgemeinerung des Hilfssatzes (6.7.1) den

(6.8.21) **Satz** *(von Taylor, 1. Fassung). Es seien* R, S *normierte Vektorräume über* $\mathbb{K}$. g *sei eine Abbildung von* $D \subset R$ *in* S; *diese sei auf der Strecke* $\overline{x\,x_0}$ n-*mal stetig differenzierbar und genüge der Ungleichung*

$$(6.8.22) \qquad |g^{(n)}(\xi) - g^{(n)}(x_0)| \leqq \gamma \cdot |\xi - x_0| \qquad (\xi \in \overline{x\,x_0}) .$$

Unter diesen Voraussetzungen ist

$$|g(x) - \sum_{\nu=0}^{n} \frac{1}{\nu!} g^{(\nu)}(x_0)\,(x - x_0)^{\nu}| \leqq \frac{\gamma}{(n+1)!} |x - x_0|^{n+1}$$

abschätzbar.

Beweis. Wir setzen wiederum o.E. voraus, daß R, S normierte Vektorräume über $\mathbb{R}$ sind. Zur Abkürzung führen wir die Abbildung

$$(6.8.23) \qquad G(\zeta) = g(\zeta) - \sum_{\nu=0}^{n} \frac{1}{\nu!} g^{(\nu)}(x_0)\,(\zeta - x_0)^{\nu}$$

ein. Der Hilfssatz (6.8.18) besagt, daß G ebenso wie g auf $\overline{x\,x_0}$ n-mal stetig differenzierbar ist, wobei insbesondere

$$(6.8.24) \qquad G(x_0) = 0,\ G'(x_0) = 0, \ldots, G^{(n)}(x_0) = 0$$

gilt. Gemäß der Folgerung (6.6.12) gibt es ein $l \in L(S, \mathbb{R})$ mit

$$l(G(x)) = |G(x)|, \quad |l| \leqq 1 .$$

Weiter setzen wir für $t \in [0, 1]$ $\varphi(t) = x_0 + t(x - x_0)$ und bilden hiermit

$$F(t) = (l \circ G \circ \varphi)(t) .$$

Wir behaupten: diese reellwertige Funktion F ist in [0, 1] n-mal stetig differenzierbar und

$$(6.8.25) \qquad F^{(\nu)}(t) = l[G^{(\nu)}(\varphi(t))\,(x - x_0)^{\nu}] \qquad (\nu = 0, 1, \ldots, n) .$$

Den Beweis dieser Aussage führen wir durch Induktion nach ν. Für $\nu = 0$ ist nichts zu zeigen. Zum Schluß von ν auf $\nu + 1$ ($\leqq$ n) kann man neben der Kettenregel (6.6.9) die Produktregel (6.6.5) heranziehen; einfacher ist es jedoch, durch direkten Grenzübergang (unter Benutzung der Kettenregel) die Existenz der Ableitung

$$F^{(\nu+1)}(t) = \{l[G^{(\nu)}(\varphi(t))\,(x - x_0)^{\nu}]\}' = l[G^{(\nu+1)}(\varphi(t))\,(x - x_0)^{\nu+1}]$$

zu erschließen.

Aus (6.8.25) folgt mit (6.8.24) speziell

$$F^{(\nu)}(0) = 0 \qquad (\nu = 0, 1, \ldots, n) .$$

Diese Gleichungen (für $\nu \leqq n-1$) verwendet man, um induktiv mittels partieller Integration für $t \in [0, 1]$ die Darstellungen

$$(6.8.26) \qquad F(t) = \int_0^t F^{(\nu)}(\tau) \frac{(t-\tau)^{\nu-1}}{(\nu-1)!}\, d\tau \qquad (\nu = 1, 2, \ldots, n)$$

herzuleiten. Nun gilt nach (6.8.25)

$$F^{(n)}(\tau) = l[G^{(n)}(\varphi(\tau))\,(x - x_0)^n] = l[(g^{(n)}(\varphi(\tau)) - g^{(n)}(x_0))\,(x - x_0)^n]\,,$$

mithin auf Grund der Voraussetzung (6.8.22) die Abschätzung

$$|F^{(n)}(\tau)| \leqq \gamma\,|\varphi(\tau) - x_0|\,|x - x_0|^n = \gamma\,|x - x_0|^{n+1}\tau\,,$$

die, auf (6.8.26) angewandt, die Behauptung

$$|G(x)| = F(1) \leqq \int_0^1 \tau(1-\tau)^{n-1} d\tau \frac{\gamma}{(n-1)!}\,|x - x_0|^{n+1} = \frac{\gamma}{(n+1)!}\,|x - x_0|^{n+1}$$

liefert.

Mit Hilfe des Mittelwertsatzes (6.6.13) folgert man den

(6.8.27) **Zusatz.** Die Ungleichung (6.8.22) ist erfüllt, falls g auf $\overline{x\,x_0}$ $(n+1)$-mal differenzierbar und

$$|g^{(n+1)}(\xi)| \leqq \gamma \qquad (\xi \in \overline{x\,x_0})$$

abschätzbar ist.

Für reellwertige Abbildungen notieren wir den in gewissem Sinne weitergehenden

(6.8.28) **Satz** *(von Taylor, 2. Fassung). Es sei* R *ein normierter Vektorraum über* $\mathbb{R}$. g *sei eine Abbildung von* $D \subset R$ *in* $\mathbb{R}$; *diese sei auf der Strecke* $\overline{x\,x_0}$ $(n+1)$*-mal differenzierbar.*

Dann existiert ein $\xi \in \overline{x\,x_0}$, $\neq x$ *und* $\neq x_0$, *so daß*

$$g(x) = \sum_{\nu=0}^{n} \frac{1}{\nu!}\, g^{(\nu)}(x_0)\,(x - x_0)^\nu + \frac{1}{(n+1)!}\, g^{(n+1)}(\xi)\,(x - x_0)^{n+1}\,.$$

Beweis. Mit $\varphi(t) = x_0 + t(x - x_0)$ betrachten wir hier die reellwertige Funktion

$$G(t) = (g \circ \varphi)(t) \qquad (t \in [0, 1])\,.$$

Gemäß den Überlegungen zum vorangehenden Satz ist G in $[0, 1]$ $(n+1)$-mal differenzierbar und dort

$$G^{(\nu)}(t) = g^{(\nu)}(\varphi(t))\,(x - x_0)^\nu \qquad (\nu = 0, 1, \ldots, n+1)\,.$$

Den erweiterten Mittelwertsatz – vgl. z.B. Erwe [14], S. 141–145 – wenden wir nun bezüglich der Funktionen

$$F(t) = \sum_{\nu=0}^{n} \frac{1}{\nu!} G^{(\nu)}(t)\,(1-t)^{\nu},\quad H(t) = (1-t)^{n+1} \qquad (t \in [0,1])$$

an; danach existiert ein $0 < \tau < 1$ mit

$$\frac{F(1)-F(0)}{H(1)-H(0)} = \frac{F'(\tau)}{H'(\tau)}.$$

Wir wählen $\xi = \varphi(\tau)$ und erhalten auf Grund der Beziehungen

$$F(1) = G(1) = g(x),\ F(0) = \sum_{\nu=0}^{n} \frac{1}{\nu!} G^{(\nu)}(0) = \sum_{\nu=0}^{n} \frac{1}{\nu!} g^{(\nu)}(x_0)\,(x-x_0)^{\nu},$$

$$H(1) = 0,\ H(0) = 1$$

sowie

$$F'(\tau) = \sum_{\nu=0}^{n} \frac{1}{\nu!} G^{(\nu+1)}(\tau)\,(1-\tau)^{\nu} - \sum_{\nu=1}^{n} \frac{1}{(\nu-1)!} G^{(\nu)}(\tau)\,(1-\tau)^{\nu-1}$$

$$= \frac{1}{n!} g^{(n+1)}(\xi)\,(x-x_0)^{n+1}(1-\tau)^{n},$$

$$H'(\tau) = -(n+1)(1-\tau)^{n}$$

durch Einsetzen unmittelbar die Behauptung.

Schwächere Voraussetzungen als bei den bisher bewiesenen *globalen* Fassungen des Satzes von Taylor genügen bei der folgenden *lokalen* Aussage.

(6.8.29) **Satz** *(von Taylor, 3. Fassung). Es seien* R, S *normierte Vektorräume über* IK. g *sei eine Abbildung von* $D \subset R$ *in* S; *diese sei in* $x_0 \in D$ n*-mal differenzierbar.*

Unter diesen Voraussetzungen gilt für $x \in D$

$$g(x) = \sum_{\nu=0}^{n} \frac{1}{\nu!} g^{(\nu)}(x_0)\,(x-x_0)^{\nu} + |x-x_0|^{n}\,\epsilon(x),$$

worin ϵ *eine in* x_0 *stetige Abbildung von* D *in* S *mit* $\epsilon(x_0) = 0$ *bezeichnet.*

Wir führen den *Beweis* durch Induktion nach n. Für $n = 1$ ist die Behauptung nach Definition (6.6.1) trivialerweise klar. Zum Schluß von n auf $n+1$ setzen wir voraus, daß g in x_0 $(n+1)$-mal differenzierbar ist; ferner können wir o.E. zusätzlich – vgl. den Beweis zu Satz (6.8.21) – von

$$g(x_0) = 0,\ g'(x_0) = 0, \ldots,\ g^{(n+1)}(x_0) = 0$$

ausgehen. Zu zeigen ist, daß

$$\epsilon(x) := \begin{cases} \dfrac{1}{|x - x_0|^{n+1}}\, g(x) & (x \in D,\ \neq x_0)\,, \\ 0 & (x = x_0) \end{cases}$$

in x_0 stetig ist.

Hierzu sei $\epsilon > 0$ vorgegeben. Wegen $n + 1 \geq 2$ ist g in x_0 mindestens zweimal differenzierbar. Demgemäß ist $x_0 \in \mathring{D}_{g'}$ und g' in x_0 n-mal differenzierbar. Die Induktionsannahme besagt, daß

$$\eta(x) := \begin{cases} \dfrac{1}{|x - x_0|^{n}}\, g'(x) & (x \in D_{g'},\ \neq x_0) \\ 0 & (x = x_0) \end{cases}$$

in x_0 stetig ist. Infolgedessen existiert ein $\rho > 0$, so daß $K(x_0, \rho) \subset D_{g'}$ sowie für $\zeta \in K(x_0, \rho)$

$$|g'(\zeta)| \leqq \epsilon\, |\zeta - x_0|^n \tag{6.8.30}$$

gilt. Es sei nun $x \in K(x_0, \rho)$. Für $\zeta \in \overline{x\, x_0}$ schätzt man nach (6.8.30)

$$|g'(\zeta) - g'(x_0)| = |g'(\zeta)| \leqq \epsilon\, |\zeta - x_0|^n \leqq (\epsilon\, |x - x_0|^{n-1})\, |\zeta - x_0|$$

ab. Bezüglich $\gamma = \epsilon\, |x - x_0|^{n-1}$ sind somit die Voraussetzungen des Hilfssatzes (6.7.1), d.h. des Satzes (6.8.21) im Spezialfall n = 1 erfüllt. Es folgt daher

$$|g(x)| = |g(x) - g(x_0) - g'(x_0)\,(x - x_0)| \leqq \frac{\gamma}{2}\, |x - x_0|^2 = \frac{\epsilon}{2}\, |x - x_0|^{n+1}\,,$$

womit der Beweis abgeschlossen ist.

Angemerkt sei, daß man

$$P(x) = \sum_{\nu=0}^{n} \frac{1}{\nu!}\, g^{(\nu)}(x_0)\,(x - x_0)^\nu$$

das n-*te oskulierende Polynom von* g *in* x_0 nennt. Speziell im Fall $R = \mathbb{R}^k$, $S = \mathbb{R}$ kann man hierin nach (6.8.12), (6.8.13) die (totale) Ableitung $g^{(\nu)}(x_0)$ durch die partiellen Ableitungen ausdrücken. Hierzu bezeichnen wir wie üblich für $\alpha = (\alpha_i)_1^k \in \mathbb{N}^k$, $x = (\xi_i)_1^k \in \mathbb{R}^k$

$$|\alpha| = \alpha_1 + \alpha_2 + \ldots + \alpha_k\,,\quad \alpha! = \alpha_1! \cdot \alpha_2! \cdot \ldots \cdot \alpha_k!\,,$$

$$\partial^\alpha = \frac{\partial^{|\alpha|}}{\partial \xi_1^{\alpha_1}\, \partial \xi_2^{\alpha_2} \ldots \partial \xi_k^{\alpha_k}}\,,\quad x^\alpha = \xi_1^{\alpha_1} \cdot \xi_2^{\alpha_2} \cdot \ldots \cdot \xi_k^{\alpha_k}$$

und erschließen unter Berücksichtigung des Satzes von Schwarz die Beziehung

$$\frac{1}{\nu!}\, g^{(\nu)}(x_0)\,(x - x_0)^\nu = \sum_{|\alpha| = \nu} \frac{1}{\alpha!}\, (\partial^\alpha g)(x_0)\,(x - x_0)^\alpha\,, \tag{6.8.31}$$

die unmittelbar die Darstellung

$$P(x) = \sum_{|\alpha|=0}^{n} \frac{1}{\alpha!} (\partial^\alpha g)(x_0)(x - x_0)^\alpha$$

impliziert; hierzu vgl. Forster [18], Bd. 2, S. 55–58.

Abschließend kommen wir kurz auf die Iterationsverfahren zurück. Dabei sei R ein normierter Vektorraum über $\mathbb{K}$, g eine Abbildung von $D \subset R$ in R, $\hat{x} \in D$ und $g(\hat{x}) = 0$.

Unter Benutzung des Satzes von Taylor leiten wir zunächst noch einmal das *Newton-Verfahren* her. Dazu sei $x_0 \in D$ eine Näherung von $\hat{x}$, g in x_0 differenzierbar und $g'(x_0) \in J(R)$. Nach Satz (6.8.29), im Trivialfall $n = 1$, gilt speziell für $x = \hat{x}$

$$0 = g(\hat{x}) = g(x_0) + g'(x_0)(\hat{x} - x_0) + |\hat{x} - x_0| \, \epsilon(\hat{x}) .$$

Liegt x_0 hinreichend nahe bei $\hat{x}$, d.h. ist $|\epsilon(\hat{x})|$ genügend klein, so läßt sich vermuten, daß x_1 in

$$0 = g(x_0) + g'(x_0)(x_1 - x_0) ,$$

d.h.

$$x_1 = x_0 - g'(x_0)^{-1} g(x_0)$$

eine bessere Näherung von $\hat{x}$ darstellt. Erfüllt x_1 die gleichen Voraussetzungen wie x_0, so definiert man in analoger Weise ein x_2 usw. Es ergibt sich so natürlich das Newton-Verfahren (6.7.5).

Wie wir in Satz (6.7.3) – siehe auch Übungsaufgabe 6.15 – festgestellt haben, ist das Newton-Verfahren *quadratisch* konvergent. Dies bedeutet, daß bei jedem Iterationsschritt die Zahl der signifikanten Dezimalstellen in etwa verdoppelt wird. Trotz der hohen Rechengeschwindigkeit der heutigen Computer ist es unter Umständen sinnvoll, ein Verfahren höherer, z.B. *kubischer* Konvergenzordnung zu verwenden; hierzu vgl. Ehrmann [13]. Zu derartigen Verfahren gelangt man folgendermaßen:

Es sei wiederum x_0 eine Näherung von $\hat{x}$, g jedoch nun zweimal differenzierbar in x_0. Abermals nach (6.8.29) hat man

$$0 = g(\hat{x}) = g(x_0) + g'(x_0)(\hat{x} - x_0) + \frac{1}{2} g''(x_0)(\hat{x} - x_0)^2 + |\hat{x} - x_0|^2 \, \epsilon(\hat{x}) .$$

Danach sollte durch

$$\begin{cases} 0 = g(x_0) + g'(x_0)\,\tilde{d}_0 + \frac{1}{2} g''(x_0)\,\tilde{d}_0^2 , \\ \tilde{x}_1 = x_0 + \tilde{d}_0 \end{cases}$$

eine in der Regel bessere Näherung $\tilde{x}_1$ bestimmt sein. Die in $\tilde{d}_0$ „quadratische Gleichung" ist aber nur schwer zu lösen, sie wird daher „linearisiert". Dies ge-

schieht auf zwei verschiedene Arten:

(i) Man berechnet der Reihe nach c_0, d_0, x_1 aus

$$(6.8.32)\quad \begin{cases} g(x_0) + g'(x_0)\, c_0 = 0\,, \\ g(x_0) + g'(x_0)\, d_0 + \frac{1}{2} g''(x_0)\, c_0\, d_0 = 0\,, \\ x_1 = x_0 + d_0 \end{cases}$$

und entsprechend rekursiv c_n, d_n, x_{n+1} aus

$$(6.8.33)\quad \begin{cases} g(x_n) + g'(x_n)\, c_n = 0\,, \\ g(x_n) + g'(x_n)\, d_n + \frac{1}{2} g''(x_n)\, c_n\, d_n = 0\,, \\ x_{n+1} = x_n + d_n\,. \end{cases}$$

Durch Einsetzen von c_n und d_n erhält man also

$$(6.8.33')\quad x_{n+1} = x_n - \left[I - \frac{1}{2} g'(x_n)^{-1} g''(x_n) g'(x_n)^{-1} g(x_n) \right]^{-1} g'(x_n)^{-1} g(x_n)\,.$$

Dabei ist natürlich vorauszusetzen, daß die auftretenden Rechenoperationen durchführbar sind. (6.8.33) bzw. (6.8.33′) nennt man das *Halley-Verfahren* oder in Anbetracht seiner geometrischen Bedeutung auch das *Verfahren der tangierenden Hyperbeln.*

(ii) Man linearisiert die zweite Gleichung in (6.8.32) weitergehender, d.h. man berechnet hier d_0 aus

$$g(x_0) + g'(x_0)\, d_0 + \frac{1}{2} g''(x_0)\, c_0^2 = 0\,.$$

Dies führt rekursiv – soweit sinnvoll – zu

$$(6.8.34)\quad \begin{cases} g(x_n) + g'(x_n)\, c_n = 0 \\ g(x_n) + g'(x_n)\, d_n + \frac{1}{2} g''(x_n)\, c_n^2 = 0\,, \\ x_{n+1} = x_n + d_n\,, \end{cases}$$

bzw. aufgelöst zu

$$(8.6.34')\quad x_{n+1} = x_n - g'(x_n)^{-1} g(x_n) - \frac{1}{2} g'(x_n)^{-1} g''(x_n) \left(g'(x_n)^{-1} g(x_n)\right)^2$$

und heißt *Tschebyscheff-Verfahren* oder *Verfahren der tangierenden Parabeln.*

In ähnlicher Weise kommt man auch zu Verfahren noch höherer Konvergenzordnung; Genaueres findet man bei Döring [9].

Während man sich zur Motivation der verschiedenen Iterationsverfahren auf die lokale Fassung des Satzes von Taylor stützt, benutzt man zum Beweis geeigneter Konvergenzsätze die globale 1. Fassung des Satzes von Taylor. Analog zum Newton-

Verfahren hat Döring Konvergenzaussagen für das Halley-Verfahren in [11] und für das Tschebyscheff-Verfahren in [12] bewiesen; dabei sind auch geeignete a priori sowie a posteriori Abschätzungen der Verfahrensfehler angegeben.

Das Halley-Verfahren sollte gemäß seiner Herleitung etwas rascher konvergieren als das Tschebyscheff-Verfahren; dies wird durch die Praxis auch voll bestätigt. Demgegenüber ist der Rechenaufwand beim Tschebyscheff-Verfahren etwas geringer.

(6.8.35) **Beispiel.** Wir berechnen die im Intervall $]0, 1[$ liegende Nullstelle des Polynoms

$$P(x) = x^4 - 5x^3 + x^2 + 1$$

mit Hilfe der erwähnten Iterationsverfahren, jeweils ausgehend von

$$x_0 = 0{,}5 \ ;$$

dabei verwenden wir eine 24-stellige Dezimalarithmetik. In den folgenden Tabellen sind die signifikanten Stellen jeder Näherung unterstrichen; hierdurch kommen die Konvergenzordnungen der Verfahren recht gut zum Vorschein.

Newton-Verfahren

k	x_k
1	0,805555555555555555555556
2	0,715417448427476286588775
3	0,703487210635426691796221
4	0,703282982376015009646237
5	0,703282922964201550252416
6	0,703282922964196523032428

Halley-Verfahren

k	x_k
1	0,681985294117647058823529
2	0,703267059230422831323212
3	0,703282922964190313975401
4	0,703282922964196523032428

Tschebyscheff-Verfahren

k	x_k
1	0,598079561042524005486969
2	0,695913747766220343656072
3	0,703281436626732652159067
4	0,703282922964196511264947
5	0,703282922964196523032428

Übungsaufgaben zum 6. Kapitel

Aufgabe 6.1. Es sei (R, d) vollständiger metrischer Raum, $D \subseteq R$, $x \in D$, T Abbildung von D in R mit $|T| < 1$. Es bezeichne

$$\rho' := \frac{1}{1-|T|}\, d(x, T(x)),$$

$$D^* := \{y \in R:\ d(y,x) \leqq \rho'\};$$

es gelte

$$D^* \subseteq D.$$

Man zeige:

(i) $T(D^*) \subseteq D^*$,

(ii) T besitzt in D genau einen Fixpunkt $\hat{x}$,

(iii) $\hat{x}$ liegt in D^*; für jedes $x_0 \in D^*$ strebt $x_n = T^n(x_0)$ gegen $\hat{x}$ für $n \to \infty$,

(iv) es gelten die Fehlerabschätzungen

$$d(x_n, x) \leqq d(x_1, x_0)\, \frac{|T|^n}{1-|T|} \qquad (n \in \mathbb{N}),$$

$$d(x_n, x) \leqq d(x_n, x_{n-1})\, \frac{|T|}{1-|T|} \qquad (n \in \mathbb{N}, > 0).$$

Anleitung: Nach der Aussage (i) beweist man die Aussage (iii), indem man den Banachschen Fixpunktsatz (6.1.7′) auf den metrischen Raum D^* anwendet; erst danach ist (ii) zu zeigen.

Aufgabe 6.2. Man berechne die kleinste positive Nullstelle von

$$f(x) = \cosh(x) - 2x$$

mit dem Iterationsverfahren

$$x_0 = 0,\ x_{n+1} = \frac{1}{2}\cosh(x_n) \quad (n = 0, 1, 2, \ldots)$$

auf 6 Dezimalstellen genau. Anschließend zeige man, daß die Voraussetzungen der Aufgabe 6.1 mit einem geeigneten D und D^* erfüllt sind und die Folge der x_n ab einem geeigneten $N \in \mathbb{N}$ in D^* liegt. Schließlich führe man eine Fehlerabschätzung durch.

Lösung: $\hat{x} = 0{,}589388$; man wählt beispielsweise $D = [0; 0{,}6]$, D^* zu $x = 0{,}5$.

Aufgabe 6.3.

(i) Man zeige: das Gleichungssystem

$$(*) \qquad \begin{cases} \xi^3 - 3\xi + 2\eta + 2 = 0, \\ \eta^2 - 6\eta - \xi - 2 = 0 \end{cases}$$

besitzt in

$$R := [0,\ 0{,}5] \times [-0{,}5,\ 0]$$

genau eine Lösung $(\hat{\xi},\ \hat{\eta})$. Hierzu forme man das Gleichungssystem $(*)$ so um, daß die erste Gleichung nur noch ξ, die zweite Gleichung nur noch η als linearen Term enthält. Auf diese Weise gelangt man zu einem Fixpunktproblem, in dem der zugehörige Operator T: $R \to R$ bezüglich der Maximums-Metrik einen Betrag $< 0{,}25$ besitzt.

(ii) Man berechne $(\hat{\xi},\ \hat{\eta})$ auf 6 Dezimalstellen genau.

(iii) Man zeige, daß das System $(*)$ in $[-1,\ 1] \times [-1,\ 1]$ außer $(\hat{\xi},\ \hat{\eta})$ keine weitere Lösung besitzt. Hierzu gebe man eine w-Norm an, bezüglich der der Operator T aus (i) in $[-1,\ 1] \times [-1,\ 1]$ kontrahierend ist.

Lösung zu (ii): $(\hat{\xi};\ \hat{\eta}) = (0{,}440210;\ -0{,}382338)$.

Aufgabe 6.4. Es sei C eine invertierbare (n,n)-Matrix, $\tilde{L}$ und $\tilde{R}$ seien wie zu (6.2.26) definiert. Bei der Ausführung von (6.2.27) werde

$$r_m = b - Cx_m,\quad x_{m+1} = x_m + d_m$$

exakt berechnet; hingegen mögen bei der Lösung der Gleichungssysteme

$$\tilde{L}f_m = Pr_m,\ \tilde{R}d_m = f_m \qquad (m = 0, 1, 2, \ldots)$$

Rundungsfehler auftreten, so daß – vgl. (3.6.27) –

$$P(C + F_m)\,d_m = Pr_m \qquad (m = 0, 1, 2, \ldots)$$

mit von m abhängigen (n,n)-Matrizen F_m gelte. Hierbei sei

$$\|C^{-1}F_m\|_w \leqq \sigma < \frac{1}{2} \qquad (m \in \mathbb{N})$$

vorausgesetzt. Man zeige, daß die Folge der x_m auch jetzt gegen die Lösung $\hat{x}$ von $Cx = b$ konvergiert.

Aufgabe 6.5. Es sei

$$G = \begin{pmatrix} 1 & 0{,}5 & 1 \\ 0{,}5 & 1 & 1 \\ -2 & 2 & 1 \end{pmatrix}.$$

Man zeige, daß das Gesamtschrittverfahren zur Lösung von $Gx = h$ konvergiert, das Einzelschrittverfahren nicht.

Aufgabe 6.6. Man zeige, daß für

$$G = \begin{pmatrix} 10 & 1 & 1 & 1 \\ 10 & 10 & 1 & 1 \\ 10 & 10 & 10 & 1 \\ 10 & 10 & 0 & 10 \end{pmatrix}$$

das Einzelschrittverfahren konvergiert, obwohl das schwache Zeilensummenkriterium und das schwache Spaltensummenkriterium nicht erfüllt sind.

Aufgabe 6.7. Man beweise Hilfssatz (6.5.11).

Aufgabe 6.8. Es sei $Q = [0, 1] \times [0, 1]$, $f(x,y) = -2\pi^2 \sin(\pi x) \sin(\pi y)$.
Man zeige:

(i) $u(x,y) := \sin(\pi x) \sin(\pi y)$
ist Lösung der Randwertaufgabe

$$\Delta u = f \text{ in } \overset{\circ}{Q}, \quad u = 0 \text{ auf } \partial Q .$$

(ii) Es sei

$$h = \frac{1}{N} \quad (N \in \mathbb{N}, > 0) .$$

Man zeige, daß die diskretisierte Aufgabe

$$\begin{cases} \Delta_h u_h(x,y) = f(x,y) & ((x,y) \in \overset{\circ}{Q}_h) , \\ u_h(x,y) = 0 & ((x,y) \in \partial Q_h) \end{cases}$$

von

$$u_h(x,y) := \frac{h^2\pi^2}{2(1-\cos\pi h)}\, u(x,y) \quad ((x,y) \in Q_h)$$

gelöst wird.

(iii) Man folgere aus (ii) für $h = \frac{1}{2M}$ $(M \in \mathbb{N}, > 0)$

$$\max \{ |u_h(x,y) - u(x,y)| : (x,y) \in Q_h \} = \frac{h^2\pi^2}{12} + O(h^4) \quad (h \to 0) .$$

Hinweis zu (ii): Man beachte, daß nach Hilfssatz (6.5.11)

$$(f(x,y))_{(x,y) \in \overset{\circ}{Q}_h}$$

ein Eigenvektor der Koeffizientenmatrix G des Gleichungssystems (6.5.7) ist.

Aufgabe 6.9. Das Gleichungssystem (6.5.7) sei als

$$Gu = c$$

geschrieben. Ausgehend von $u_0 = 0$, sei $(u_m)_0^\infty$ die mit dem Gesamtschrittverfahren konstruierte Folge von Näherungslösungen; es bezeichne

$$r_m := c - Gu_m \quad (m \in \mathbb{N}) .$$

Man zeige:

(i) $r_m = 4 A_{Ges}(u_m - u_{m-1})$,

(ii) $u_m - u_{m-1} = A_{Ges}^{m-1}(u_1 - u_0) = \frac{1}{4} A_{Ges}^{m-1} c$.

In der Zerlegung von c in Eigenvektoren zu A_{Ges}, also

$$c = \sum_{p,q=1}^{N-1}{}' \alpha_{p,q} x_{p,q} \quad (\alpha_{p,q} \in \mathbb{R})$$

mit den $x_{p,q}$ aus (6.5.11) sei

$$|\alpha_{1,1}|^2 + |\alpha_{N-1,N-1}|^2 \neq 0$$

vorausgesetzt. Wir bezeichnen

$$\rho = \begin{cases} 0 & \text{im Fall } N \leqq 2, \\ \frac{1}{2}\left(1 + \dfrac{\cos(\frac{2\pi}{N})}{\cos(\frac{\pi}{N})}\right) & (<1) \text{ im Fall } N \geqq 3. \end{cases}$$

Man folgere mit den Methoden von Kapitel 5.1 aus (ii) folgendes Verhalten bezüglich der euklidischen Norm:

(iii) $|A_{Ges}(u_m - u_{m-1})| = \cos\left(\frac{\pi}{N}\right)|u_m - u_{m-1}|(1 + O(\rho^m)) \quad (m \to \infty)$.

Hieraus ergibt sich unter Berücksichtigung von (i) die asymptotische Übereinstimmung der Abschätzungen (6.5.19) und (6.5.22).

Aufgabe 6.10.

(i) Man beweise die Produktregel, also Satz (6.6.5).

(ii) Es sei zusätzlich $\tilde{f}: D \to S_2$, differenzierbar in x_0 gegeben. Man zeige: die Abbildung

$$\tilde{h}: D \to S_3$$

mit

$$\tilde{h}(x) = g(x)\tilde{f}(x)$$

ist differenzierbar in x_0, und man hat für $k \in R$

$$\tilde{h}'(x_0)k = (g'(x_0)k)\tilde{f}(x_0) + g(x_0)(\tilde{f}'(x_0)k).$$

Zum Beweis verwende man (i) und beachte die Normisomorphie

$$S_2 \simeq L(\mathbb{K}, S_2).$$

Aufgabe 6.11. Mit Hilfe des Fixpunktsatzes (6.1.7) beweise man die folgende Version des Satzes über das Newton-Verfahren:

Es sei R Banach-Raum über $\mathbb{K}$, $x_0 \in R$, $\rho > 0$ und $\kappa \geqq 0$; es bezeichne

$$K(x_0, \rho) := \{x \in R: |x - x_0| \leqq \rho\}.$$

Es sei

$$g: K(x_0, \rho) \to R$$

zweimal differenzierbar; ferner gelte für alle $x \in K(x_0, \rho)$

$$|g''(x)| \leqq \kappa ,$$
$$g'(x) \in J(R),\ |g'(x)^{-1}| \cdot |g'(x)^{-1} g(x)| \cdot \kappa \leqq p < 1$$

und schließlich

$$\frac{1}{1-p} |g'(x_0)^{-1} g(x_0)| \leqq \rho .$$

Dann gilt:

(i) g besitzt in $K(x_0, \rho)$ genau eine Nullstelle $\hat{x}$;

(ii) die durch

$$x_{n+1} = x_n - g'(x_n)^{-1} g(x_n) \quad (n \in \mathbb{N})$$

definierte Folge strebt für $n \to \infty$ gegen $\hat{x}$,

(iii) es gelten die Abschätzungen

$$|x_n - \hat{x}| \leqq \frac{p^n}{1-p} |x_1 - x_0| \qquad (n \in \mathbb{N}) ,$$

$$|x_n - \hat{x}| \leqq \frac{p}{1-p} |x_n - x_{n-1}| \qquad (n \in \mathbb{N}, > 0) .$$

Aufgabe 6.12. Es sei R Banach-Raum, $\hat{x} \in R$, $r' > 0$, $\kappa > 0$. Es sei g eine zweimal differenzierbare Abbildung von $K(x, r') \subseteq R$ in R mit den Eigenschaften

$$g(\hat{x}) = 0,\ g'(\hat{x}) \in J(R)$$

und

$$|g''(x)| \leqq \kappa \quad (x \in K(\hat{x}, r')) ;$$

hierzu bezeichnen wir

$$\omega := |g'(\hat{x})^{-1}| .$$

Ist dann

$$r = \min \left\{ \frac{1}{3} r',\ \frac{1}{12} (\omega\kappa)^{-1} \right\} ,$$

so erfüllen alle $x_0 \in R$ mit $|x_0 - \hat{x}| \leqq r$ die Voraussetzungen der Aufgabe 6.11, und zwar mit

$$\rho = 2r .$$

Aufgabe 6.13. Es seien die Voraussetzungen des Satzes (6.7.3) über das Newton-Verfahren und zusätzlich die schärfere Bedingung

$$h \leqq \frac{1}{4}$$

erfüllt. Man zeige, daß auch das vereinfachte Newton-Verfahren mit

$$A = g'(x_0)$$

gegen $\hat{x}$ konvergiert. Bei der Anwendung des Satzes (6.6.14) wähle man

$$\rho = 2\,|g'(x_0)^{-1}\,g(x_0)|$$

und beachte, daß statt der Bedingung

$$\sigma \leqq \mu\rho(1-p)$$

die schwächere Voraussetzung

$$|A^{-1}\,g(x_0)| \leqq \rho(1-p)$$

genügt.

Da das vereinfachte Newton-Verfahren für x_1 den gleichen Wert wie das Newton-Verfahren liefert, läßt sich die Fehlerabschätzung (6.6.16) für $|x_1 - \hat{x}|$ auch im Newton-Verfahren anwenden.

Aufgabe 6.14. Es seien die Voraussetzungen der Ergänzung (6.7.30) erfüllt und hierbei

$$r := \frac{1}{4}\min\{r', (\omega\kappa)^{-1}\}.$$

Man zeige: g besitzt in $K(\hat{x}, 4r)$ nur die Nullstelle $\hat{x}$.

Anleitung: Für $x \in K(\hat{x}, 4r)$ zeige man unter Benutzung von (6.7.1) die Ungleichung

$$|g'(\hat{x})^{-1}\,g(x)| \geqq \frac{1}{2}\,|x - \hat{x}|\,.$$

Aufgabe 6.15. Es sei R Banach-Raum, $\hat{x} \in R$, $(x_n)_0^\infty$ eine Folge in R. Man sagt, die Folge $(x_n)_0^\infty$ konvergiert *linear* oder *von 1. Ordnung* gegen $\hat{x}$ genau dann, wenn ein p mit $0 < p < 1$ existiert, so daß für alle n

$$|x_{n+1} - \hat{x}| \leqq p\,|x_n - \hat{x}|$$

gilt. Für $k \geqq 2$ heißt die Folge $(x_n)_0^\infty$ *von k-ter Ordnung konvergent gegen* $\hat{x}$ genau dann, wenn x_n gegen $\hat{x}$ konvergiert und ein $c > 0$ existiert, so daß für alle $n \in \mathbb{N}$

$$|x_{n+1} - \hat{x}| \leqq c\,|x_n - \hat{x}|^k\,.$$

Im Fall $k = 2$ spricht man auch von *quadratischer*, im Fall $k = 3$ von *kubischer* Konvergenz.

Man zeige:

(i) Die Folge des vereinfachten Newton-Verfahrens in Satz (6.6.14) konvergiert linear gegen $\hat{x}$.

(ii) Unter der zusätzlichen Voraussetzung

$$g'(\hat{x}) \in J(R)$$

konvergiert in Satz (6.7.3) die Folge des Newton-Verfahrens quadratisch gegen $\hat{x}$. Hierzu zeige man, daß die Folge der $|g'(x_n)^{-1}|$ beschränkt ist, und benutze die Beziehungen

$$x_{n+1} - \hat{x} = g'(x_n)^{-1}[g'(x_n)(x_n - \hat{x}) - (g(x_n) - g(\hat{x}))] .$$

(iii) Für das Beispiel

$$g(x) = x^2 \quad (g: \mathbb{R} \to \mathbb{R}), \quad x_0 \in \mathbb{R}, \neq 0 \text{ beliebig}$$

zeige man, daß die Voraussetzungen des Satzes (6.7.3) erfüllt sind, jedoch das Newton-Verfahren nur linear konvergiert.

Aufgabe 6.16. Es sei $D \subset \mathbb{R}^n$, g eine Abbildung von D in $\mathbb{R}^n$, ferner $x_0 \in \mathring{D}$, g differenzierbar in x_0. Man zeige:

(i) Die Abbildung h: $D \to \mathbb{R}$, definiert durch

$$h(x) := \|g(x)\|_2^2 = g(x)^t g(x)$$

ist in x_0 differenzierbar, für $k \in \mathbb{R}^n$ gilt

$$h'(x_0)k = 2\, g(x_0)^t g'(x_0)k .$$

(ii) Es sei zusätzlich $g(x_0) \neq 0$, $g'(x_0)$ invertierbar. Dann existiert ein $0 < \lambda_0 < 1$, so daß für

$$\begin{cases} 0 < \lambda \leqq \lambda_0 , \\ x_\lambda := x_0 - \lambda\, g'(x_0)^{-1} g(x_0) \end{cases}$$

gilt

$$x_\lambda \in D, \quad h(x_\lambda) \leqq (1 - \lambda)\, h(x_0) < h(x_0) .$$

Aufgabe 6.17. Vorgegeben sei die Eigenwertaufgabe

$$Ay = \lambda By$$

mit

$$A = B = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} .$$

(i) Man zeige, daß für alle $\lambda \in \mathbb{C}, \neq 1$

$$\dim H_\lambda(A, B) = 1 .$$

(ii) Es sei g(x) gemäß (6.7.35) definiert, und zwar mit

$$s(x) = s(y) := \eta_1 \quad (y = (\eta_i)_1^2 \in \mathbb{C}^2) .$$

Ferner sei $\hat{\lambda} \in \mathbb{C}, \neq 1$ vorgegeben und hierzu

$$\hat{x} := \begin{pmatrix} e_1 \\ \hat{\lambda} \end{pmatrix} \in \mathbb{C}^2 \times \mathbb{C} .$$

Man zeige:

$$\begin{cases} g(\hat{x}) = 0 , \\ g'(\hat{x}) \text{ ist nicht invertierbar.} \end{cases}$$

Aufgabe 6.18. Es sei $g(x)$ gemäß (6.7.35) definiert und dabei

$$s(x) = s(y) := \eta_{i_0} \quad (y = (\eta_i)_1^n \in \mathbb{C}^n) .$$

Bei Anwendung des Newton-Verfahrens auf die Abbildung g berechnet man, von einem $x_0 \in \mathbb{C}^n \times \mathbb{C}$ ausgehend, für $k = 0, 1, 2, \ldots$

$$\begin{cases} g'(x_k) r_k = g(x_k) , \\ x_{k+1} = x_k - r_k . \end{cases}$$

Es sei bezeichnet

$$r_k = \begin{pmatrix} d_k \\ \rho^{(k)} \end{pmatrix} , \quad x_k = \begin{pmatrix} y_k \\ \lambda^{(k)} \end{pmatrix} , \quad d_k = (\delta_i^{(k)})_{i=1}^n , \quad y_k = (\eta_i^{(k)})_{i=1}^n .$$

Man zeige:

(i) $g'(x_k)$ entspricht der $(n+1, n+1)$-Matrix

$$\left(\begin{array}{c|c} A - \lambda^{(k)} B & -By_k \\ \hline 0 \ldots 0 \; \underset{\uparrow}{1} \; 0 \ldots 0 & 0 \end{array} \right)$$

$$\qquad\qquad\quad i_0$$

(ii) Unter der Voraussetzung

$$\eta_{i_0}^{(0)} = 1$$

gilt für alle k

$$\delta_{i_0}^{(k)} = 0, \quad \eta_{i_0}^{(k)} = 1 ;$$

ferner genügen die Komponenten $\delta_i^{(k)}$ $(i \neq i_0)$ von r_k einem Gleichungssystem, das aus

$$g'(x_k) r_k = g(x_k)$$

durch Streichen der letzten Zeile und i_0-ten Spalte entsteht.

Aufgabe 6.19. Es seien $A, B \in M(n \times n, \mathbb{C})$, $s: \mathbb{C}^n \times \mathbb{C} \to \mathbb{C}$ linear und hierzu $g: \mathbb{C}^n \times \mathbb{C} \to \mathbb{C}^n \times \mathbb{C}$ durch (6.7.35) gegeben. Man zeige:

(i) g ist in $\mathbb{C}^n \times \mathbb{C}$ zweimal differenzierbar; man hat für

$$x = \begin{pmatrix} y \\ \lambda \end{pmatrix}, \quad h = \begin{pmatrix} c \\ \epsilon \end{pmatrix}, \quad k = \begin{pmatrix} d \\ \delta \end{pmatrix} \in \mathbb{C}^n \times \mathbb{C}:$$

$$g''(x)hk = \begin{pmatrix} -\delta Bc - \epsilon Bd \\ 0 \end{pmatrix} .$$

(ii) Bezüglich der Norm

$$|x| = \max \{|y|, |\lambda|\} \qquad \left(x = \begin{pmatrix} y \\ \lambda \end{pmatrix} \in \mathbb{C}^n \times \mathbb{C}\right)$$

gilt

$$|g''(x)| \leqq 2|B| .$$

(iii) g ist in $\mathbb{C}^n \times \mathbb{C}$ sogar beliebig oft differenzierbar mit den Ableitungen

$$g^{(\nu)}(x) = 0 \in L_\nu(\mathbb{C}^n \times \mathbb{C}, \mathbb{C}^n \times \mathbb{C}) \quad (\nu \geqq 3) .$$

7. Interpolation

In diesem Kapitel beschäftigen wir uns mit Problemen der folgenden Gestalt: Vorgegeben ist eine Klasse F von Funktionen, ferner $n+1$ verschiedene *Stützstellen* $x_0, x_1, \ldots, x_n$ sowie $n+1$ *Stützwerte* $f_0, f_1, \ldots, f_n$ in $\mathbb{R}$ bzw. $\mathbb{C}$. Gesucht ist eine Funktion $P \in F$ mit der Eigenschaft

$$P(x_i) = f_i \quad (i = 0, 1, \ldots, n) .$$

Ist hierbei F die Gesamtheit der Polynome vom Grad $\leqq n$, so spricht man von *Polynom-Interpolation;* darüber hinaus befassen wir uns mit der Interpolation durch trigonometrische, rationale und Spline-Funktionen.

Die Interpolationsaufgabe kann dahingehend verallgemeinert werden, daß in den Stützstellen auch Ableitungen bis zu einer gewissen Ordnung vorgegeben sind: diese *Hermitesche* Interpolation betrachten wir allerdings nur für den Fall der Polynome.

Die Interpolation ist die theoretische Grundlage für zahlreiche weitergehende numerische Verfahren, so beispielsweise für die numerische Quadratur und in beschränktem Umfange auch für die numerische Lösung gewöhnlicher Differentialgleichungen, wie sie in Band 3 behandelt werden. Dabei wird eine in der Regel kompliziertere Funktion f mittels der Stützwerte $f_i = f(x_i)$ durch eine numerisch zugänglichere Funktion $P \in F$ ersetzt.

Beim ursprünglichen Problem der Interpolation ging es darum, eine tabellierte Funktion f für solche x, die zwischen den Tafelstützstellen liegen, näherungsweise zu berechnen. Diese Aufgabenstellung hat jedoch in den letzten Jahren stark an Bedeutung verloren, da in den modernen Rechenmaschinen die wichtigsten transzendenten bzw. elementaren Speziellen Funktionen fest verdrahtet sind; man benutzt Funktionstafeln fast nur noch bei sehr komplizierten bzw. selten auftretenden Funktionen.

7.1. Polynom-Interpolation

Vorgegeben seien $x_0, x_1, \ldots, x_n \in \mathbb{C}$, und zwar paarweise verschieden, sowie $f_0, f_1, \ldots, f_n \in \mathbb{C}$. Gesucht ist ein Polynom P mit

$$(7.1.1) \qquad \begin{cases} \text{Grad}(P) \leqq n , \\ P(x_i) = f_i \quad (i = 0, 1, \ldots, n) . \end{cases}$$

Hierzu beweisen wir den

(7.1.2) **Satz.** *Die Newtonsche Interpolationsaufgabe* (7.1.1) *besitzt genau eine Lösung.*

Beweis.

(i) Um die Eindeutigkeit nachzuweisen, betrachten wir Polynome P, Q vom Grad $\leqq n$ mit

$$P(x_i) = Q(x_i) = f_i \quad (i = 0, 1, \ldots, n)$$

und hiermit $S = P - Q$. Offensichtlich hat S einen Grad $\leqq n$ sowie mindestens $n + 1$ Nullstellen – es ist nämlich

$$S(x_i) = P(x_i) - Q(x_i) = 0 \quad (i = 0, 1, \ldots, n)$$

–, woraus bekanntlich $S = 0$, d.h. $P = Q$ folgt.

(ii) Die Existenz einer (der) Lösung zeigen wir, indem wir diese explizit angeben. Wir bezeichnen

$$(7.1.3) \qquad p_k(x) := \prod_{\kappa=0}^{k-1} (x - x_\kappa) \quad (k = 0, 1, \ldots, n+1)$$

und definieren hiermit die *elementaren Lagrange-Polynome* zu den Stützstellen $x_0, x_1, \ldots, x_n$ durch

$$l_j^n(x) := \frac{p_{n+1}(x)}{(x - x_j)\, p'_{n+1}(x_j)} \qquad (j = 0, 1, \ldots, n).$$

Wegen

$$p'_{n+1}(x_j) = \prod_{\substack{\kappa=0 \\ \kappa \neq j}}^{n} (x_j - x_\kappa)$$

besitzt l_j^n offenbar die Darstellung

$$(7.1.4) \qquad l_j^n(x) = \prod_{\substack{\kappa=0 \\ \kappa \neq j}}^{n} \frac{x - x_\kappa}{x_j - x_\kappa} \quad (x \in \mathbb{C}).$$

Die l_j^n sind demnach Polynome vom Grad n und haben an den Stützstellen die Werte

$$l_j^n(x_i) = \begin{cases} 0 & (i \neq j), \\ 1 & (i = j). \end{cases}$$

Demgemäß erfüllt das Polynom

$$(7.1.5) \qquad P(x) = \sum_{j=0}^{n}{}' f_j \cdot l_j^n(x)$$

die Bedingung (7.1.1), womit der Existenz- und Eindeutigkeitssatz (7.1.2) bewiesen ist.

Man nennt (7.1.5) die *Lagrange-Interpolationsformel* oder die *Lagrange-Darstellung* der Lösung zu (7.1.1). Diese Darstellung ist wichtig für eine Reihe weiterführender Überlegungen. Bei der numerischen Anwendung erweist sich jedoch die Berechnung der $l_j^n(x)$ als recht mühsam, so daß man die Formel (7.1.5) nur bei kleinem n oder dann, wenn mehrere Funktionen bei gleichen Stützstellen und gleichem x zu interpolieren sind, benutzt.

Numerisch vorteilhafter sind im allgemeinen die *Algorithmen von Neville und Aitken.* Dabei wird das Interpolationspolynom P schrittweise aufgebaut. Eingehender darstellen werden wir hier das Nevillesche Verfahren; bezüglich der Variante von Aitken sei auf die Übungsaufgabe 7.1 verwiesen. – Es bezeichne für $0 \leqq s \leqq s+\mu \leqq n$ P_s^μ das durch die Bedingungen

$$(7.1.6) \quad \begin{cases} \operatorname{Grad}(P_s^\mu) \leqq \mu\,, \\ P_s^\mu(x_i) = f_i \qquad (i = s, s+1, \ldots, s+\mu) \end{cases}$$

eindeutig bestimmte Interpolationspolynom. Grundlegend ist der

(7.1.7) **Satz** *(Nevillesche Formel). Für* $x \in \mathbb{C}$ *gilt*

$$P_s^0(x) = f_s \quad (s = 0, 1, \ldots, n)$$

sowie

$$(7.1.8) \quad P_s^{\mu+1}(x) = \frac{(x - x_s)\,P_{s+1}^\mu(x) - (x - x_{s+\mu+1})\,P_s^\mu(x)}{x_{s+\mu+1} - x_s}$$

$$(s = 0, 1, \ldots, n-\mu-1;\ \mu = 0, 1, \ldots, n-1)\,.$$

Beweis. Die Beziehungen $P_s^0(x) = f_s$ sind klar; zur Begründung von (7.1.8) bezeichnen wir das auf der rechten Seite der Gleichung stehende (von s, μ abhängige) Polynom mit Q. Wegen

$$\operatorname{Grad}(P_{s+1}^\mu) \leqq \mu, \quad \operatorname{Grad}(P_s^\mu) \leqq \mu$$

ist zunächst

(*) $\operatorname{Grad}(Q) \leqq \mu + 1\,.$

Weiterhin hat man für $i = s+1, \ldots, s+\mu$

$$P_{s+1}^\mu(x_i) = P_s^\mu(x_i) = f_i$$

und daher

$$Q(x_i) = \frac{(x_i - x_s) - (x_i - x_{s+\mu+1})}{x_{s+\mu+1} - x_s}\, f_i = f_i\,.$$

Ferner gilt

$$Q(x_s) = -\frac{x_s - x_{s+\mu+1}}{x_{s+\mu+1} - x_s}\, P_s^\mu(x_s) = f_s\,,$$

ebenso erhält man

$$Q(x_{s+\mu+1}) = f_{s+\mu+1}\,,$$

insgesamt also

(**) $Q(x_i) = f_i \quad (i = s, s+1, \ldots, s+\mu+1)\,.$

Gemäß Definition löst $P_s^{\mu+1}$ die Interpolationsaufgabe (*), (**), womit die Behauptung

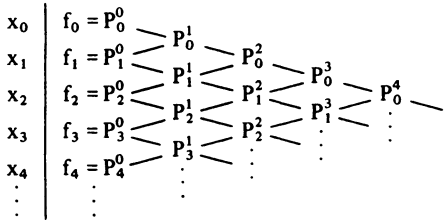
$$Q = P_s^{\mu+1}$$

bewiesen ist.

Das *Nevillesche Verfahren* besteht nun darin, bei vorgegebenen Stützstellen x_i sowie Stützwerten f_i und festem $x \in \mathbb{C}$ über die Rekursionen in Satz (7.1.7) $P(x) = P_0^n(x)$ zu berechnen. Dazu verwendet man in naheliegender Weise das folgende Schema, worin zur Abkürzung

$$P_s^\mu := P_s^\mu(x)$$

gesetzt ist:



Es interessiert natürlich insbesondere der Fall

$$x \neq x_i \quad (i = 0, 1, \ldots, n):$$

Hier kann man (7.1.8) gemäß

$$P_s^{\mu+1}(x) = \frac{(x_{s+\mu+1} - x_s)\, P_{s+1}^\mu(x) + (x - x_{s+\mu+1})\,(P_{s+1}^\mu(x) - P_s^\mu(x))}{x_{s+\mu+1} - x_s}$$

$$= P_{s+1}^\mu(x) + \frac{P_{s+1}^\mu(x) - P_s^\mu(x)}{\dfrac{x_{s+\mu+1} - x_s}{x - x_{s+\mu+1}}}$$

umformen, woraus sich unmittelbar

$$(7.1.9) \qquad P_s^{\mu+1}(x) = P_{s+1}^\mu(x) + \frac{P_s^\mu(x) - P_{s+1}^\mu(x)}{1 - \dfrac{x - x_s}{x - x_{s+\mu+1}}}$$

ergibt. Diese Formel ist vor allem dann vorteilhaft, wenn sich $P_s^{\mu+1}(x)$ nur wenig von $P_{s+1}^\mu(x)$ unterscheidet: unter dieser Voraussetzung wird nämlich als letzte Rechenoperation in (7.1.9) zu $P_{s+1}^\mu(x)$ eine Zahl kleinen Betrages addiert, was bekanntlich – vgl. (1.3.8) – zur Fehlerdämpfung führt.

Die Algorithmen von Neville und Aitken sind weniger brauchbar, sobald eine koeffizientenweise Darstellung des Interpolationspolynoms P erwünscht ist, wie

z.B. dann, wenn mehrere Funktionswerte von P berechnet werden müssen. Unter diesem Aspekt ist die *Newtonsche Darstellung* des Interpolationspolynoms vorzuziehen; ihre Herleitung bereiten wir vor mit der

(7.1.10) **Definition.** Die *dividierten Differenzen* k-*ter Ordnung*

$$\Delta f[x_i, x_{i+1}, \ldots, x_{i+k}] \quad (0 \leqq i \leqq i+k \leqq n)$$

sind rekursiv folgendermaßen erklärt:

$$\begin{cases} \Delta f[x_i] := f_i \quad (0 \leqq i \leqq n)\,, \\ \Delta f[x_i, x_{i+1}, \ldots, x_{i+k+1}] := \dfrac{\Delta f[x_{i+1}, \ldots, x_{i+k+1}] - \Delta f[x_i, \ldots, x_{i+k}]}{x_{i+k+1} - x_i} \\ \qquad (0 \leqq i \leqq i+k \leqq n-1)\,. \end{cases}$$

Hiermit notieren wir den

(7.1.11) **Satz** *(Newtonsche Interpolationsformel). Das Interpolationspolynom* $P = P_0^n$ *zu* (7.1.1) *besitzt mit den Bezeichnungen* (7.1.3), (7.1.10) *die Darstellung*

$$P(x) = \sum_{k=0}^{n} \Delta f[x_0, x_1, \ldots, x_k] \cdot p_k(x)\,.$$

Wir *beweisen* etwas allgemeiner die Formel

$$(7.1.12) \qquad P_s^{\mu}(x) = \sum_{k=0}^{\mu} \Delta f[x_s, x_{s+1}, \ldots, x_{s+k}] \prod_{\kappa=0}^{k-1} (x - x_{s+\kappa})\,,$$

und zwar durch Induktion über μ. Für $\mu = 0$ ist

$$P_s^0(x) = f_s = \Delta f[x_s]\,.$$

Zum Schluß von μ auf $\mu + 1$ betrachten wir das Polynom

$$Q(x) = P_s^{\mu+1}(x) - P_s^{\mu}(x)\,;$$

dieses besitzt offensichtlich einen Grad $\leqq \mu + 1$ sowie die $\mu + 1$ Nullstellen

$$Q(x_i) = P_s^{\mu+1}(x_i) - P_s^{\mu}(x_i) = 0 \quad (i = s, s+1, \ldots, s+\mu)\,.$$

Folglich existiert ein $c_{\mu+1} \in \mathbb{C}$, so daß

$$(7.1.13) \qquad P_s^{\mu+1}(x) = P_s^{\mu}(x) + c_{\mu+1} \prod_{\kappa=0}^{\mu} (x - x_{s+\kappa})$$

gilt. Zu bestimmen bleibt $c_{\mu+1}$. Nach Induktionsannahme hat man

$$P_s^{\mu}(x) = \Delta f[x_s, x_{s+1}, \ldots, x_{s+\mu}]\, x^{\mu} + \ldots\,,$$

$$P_{s+1}^{\mu}(x) = \Delta f[x_{s+1}, x_{s+2}, \ldots, x_{s+\mu+1}]\, x^{\mu} + \ldots\,,$$

mithin gemäß (7.1.8)

$$P_s^{\mu+1}(x) = \frac{\Delta f[x_{s+1}, \ldots, x_{s+\mu+1}] - \Delta f[x_s, \ldots, x_{s+\mu}]}{x_{s+\mu+1} - x_s} x^{\mu+1} + \ldots ,$$

woran sich wegen (7.1.10) durch Koeffizientenvergleich mit (7.1.13), wie behauptet,

$$c_{\mu+1} = \Delta f[x_s, x_{s+1}, \ldots, x_{s+\mu+1}]$$

ablesen läßt.

Wir vermerken als

(7.1.14) **Zusatz.** Die dividierten Differenzen sind symmetrisch in ihren Argumenten, d.h. für jede Permutation $(s_0, s_1, \ldots, s_\mu)$ der Zahlen $(s, s+1, \ldots, s+\mu)$ gilt

$$\Delta f[x_{s_0}, x_{s_1}, \ldots, x_{s_\mu}] = \Delta f[x_s, x_{s+1}, \ldots, x_{s+\mu}] .$$

Beweis. Nach (7.1.12) ist $\Delta f[x_s, x_{s+1}, \ldots, x_{s+\mu}]$ der Koeffizient von x^μ im Polynom P_s^μ. Dieses ist durch die Eigenschaften (7.1.6) eindeutig bestimmt, insbesondere unabhängig davon, in welcher Reihenfolge die x_i $(s \leqq i \leqq s+\mu)$ aufgeführt sind.

Nun zur numerischen Anwendung der Newtonschen Interpolationsformel: Zunächst berechnet man die dividierten Differenzen gemäß der Rekursion in (7.1.10) hierzu verwendet man das

(7.1.15) **Schema.**

x_0	$f_0 = \Delta f[x_0]$			
		$\Delta f[x_0, x_1]$		
x_1	$f_1 = \Delta f[x_1]$		$\Delta f[x_0, x_1, x_2]$	
		$\Delta f[x_1, x_2]$		$\Delta f[x_0, x_1, x_2, x_3]$
x_2	$f_2 = \Delta f[x_2]$		$\Delta f[x_1, x_2, x_3]$	
		$\Delta f[x_2, x_3]$		
x_3	$f_3 = \Delta f[x_3]$			
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Sodann ermittelt man für ein vorgegebenes $\alpha \in \mathbb{C}$ den Wert $P(\alpha)$ mit Hilfe des

(7.1.16) *Algorithmus.*

$$c_k := \Delta f[x_0, x_1, \ldots, x_k] \quad (k = 0, 1, \ldots, n),$$
$$s_n := c_n ,$$
$$s_{k-1} := c_{k-1} + (\alpha - x_{k-1})\, s_k \quad (k = n, n-1, \ldots, 1),$$
$$P(\alpha) := s_0 ;$$

dabei stehen die für $0 \leqq k \leqq n$ benötigten $\Delta f[x_0, x_1, \ldots, x_k]$ in der obersten Schrägzeile des Schemas (7.1.15).

Dieses Verfahren wird in der Übungsaufgabe 7.2 begründet; zusätzlich wird dort gezeigt, daß die Formeln (7.1.16) ähnliche Anwendungsmöglichkeiten wie der Horner-Algorithmus (vgl. (1.4.2), (1.4.4)!) zulassen.

Die letzte bei der Anwendung von (7.1.16) durchzuführende Rechenoperation ist die Addition

$$P(\alpha) = c_0 + (\alpha - x_0) \cdot s_1 = f_0 + (\alpha - x_0) \cdot s_1 \ .$$

Ist $f_0 \neq 0$ und liegt α hinreichend nahe bei x_0, so daß $|(\alpha - x_0) \cdot s_1|$ wesentlich kleiner als $|f_0|$ ist, so tritt hierbei der numerisch günstige Fall der Fehlerdämpfung – vgl. (1.3.8)! – ein. Aus diesem Grund sollten die Stützstellen $x_0, x_1, \ldots, x_n$ so numeriert sein, daß das gegebene α möglichst nahe bei x_0 liegt.

Wichtig ist dabei, daß eine hierzu notwendige Umnumerierung der Stützstellen keine Neuberechnung der zugehörigen dividierten Differenzen nötig macht. Zu jedem $i_0 \in \{1, 2, \ldots, n\}$ gibt es nämlich auf Grund der Symmetrieeigenschaft (7.1.14) mindestens eine Permutation $(i_0, i_1, \ldots, i_n)$ von $(0, 1, \ldots, n)$, bezüglich der sämtliche

$$\Delta f[x_{i_0}, \ldots, x_{i_k}] \quad (k = 0, 1, \ldots, n)$$

in (7.1.15) aufgeführt sind. Beispielsweise kann man die Permutation

$$(i_0, i_0 - 1, \ldots, 0, i_0 + 1, \ldots, n)$$

wählen; die zugehörigen dividierten Differenzen

$$\Delta f[x_{i_0}] \ ,$$
$$\Delta f[x_{i_0}, x_{i_0-1}] = \Delta f[x_{i_0-1}, x_{i_0}] \ ,$$
$$\vdots$$
$$\Delta f[x_{i_0}, x_{i_0-1}, \ldots, x_0] = \Delta f[x_0, \ldots, x_{i_0-1}, x_{i_0}]$$
$$\vdots$$

sind dann offenbar sämtlich in (7.1.15) angegeben.

Bei den folgenden Überlegungen gehen wir davon aus, daß eine vorgegebene Funktion f interpoliert wird; es interessiert dann eine Darstellung und Abschätzung des hierbei entstehenden Fehlers.

(7.1.17) **Satz.** *Es sei* f *eine Abbildung von* $D \subset \mathbb{C}$ *in* $\mathbb{C}$. *Ferner seien* $x_0, x_1, \ldots, x_n \in D$, *paarweise verschieden, sowie*

$$f_i = f(x_i) \quad (i = 0, 1, \ldots, n) \ ;$$

P *sei das zugehörige Interpolationspolynom.*

Dann gilt für $x \in D \setminus \{x_0, x_1, \ldots, x_n\}$ *im Sinne der Bezeichnungen* (7.1.3), (7.1.10) *die Restglieddarstellung*

(7.1.18) $$f(x) - P(x) = \Delta f[x_0, x_1, \ldots, x_n, x] \cdot p_{n+1}(x) \ ,$$

wobei die dividierte Differenz $\Delta f[x_0, x_1, \ldots, x_n, x]$ *natürlich bezüglich*

(7.1.19) $$(x_0, f_0), (x_1, f_1), \ldots, (x_n, f_n), (x_{n+1}, f_{n+1}) := (x, f(x))$$

zu bilden ist.

«Kurz folgendes . . .

. . . Gib die Poesie zu vierhundert Gulden her, eins der Kinderbilder zu siebenhundert Gulden, die Iphigenie zu zweitausend Frank. – Da mein Talent mein Kapital ist, so bin ich gezwungen, mir Ruhe zu schaffen. – Da ich voraussichtlich von Dir nichts zu erwarten habe, ich aber zu Ende des Monats notwendig zur Arbeit kommen muß, sonst verstört sich mein Geist, so werde ich versuchen, anderwärts hundert Taler auf drei Monate aufzunehmen . . . Ich gebe die Pietà zu dreitausend Gulden, im schlimmsten Fall zu zweitausendfünfhundert Gulden, weniger geht gegen die Ehre. – Ich bin arm und verschuldet, und jeder Lump, dem ich unter weniger wahnsinnigen Verhältnissen einen Tritt geben würde, darf meine Ehre angreifen und über mich sprechen.»

So schrieb Anselm Feuerbach am 15. August 1863 an seine Mutter. Wer kein Geld hat, der hat auch keinen Mut. Das hatte 75 Jahre vorher der Freiherr Knigge geschrieben.

Beweis. Es bezeichne Q das Interpolationspolynom zu den in (7.1.19) angegebene Stützstellen bzw. -werten. Gemäß Satz (7.1.11) gilt für $z \in \mathbb{C}$

$$Q(z) = P(z) + \Delta f[x_0, x_1, \ldots, x_n, x] \cdot p_{n+1}(z) ,$$

woraus speziell für $z = x$

$$f(x) - P(x) = Q(x) - P(x) = \Delta f[x_0, x_1, \ldots, x_n, x] \cdot p_{n+1}(x) ,$$

also die Behauptung des Satzes folgt.

Im reellen Fall, der für die Anwendungen besonders wichtig ist, können wir weitere Restglieddarstellungen angeben.

(7.1.20) **Satz.** *Es sei* f *eine* $(n+1)$*-mal differenzierbare Abbildung von* $[a,b] \subset \mathbb{R}$ *in* $\mathbb{R}$. *Ferner sei (ohne Einschränkung)*

$$a \leqq x_0 < x_1 < \ldots < x_n \leqq b, \quad f_i = f(x_i) \quad (i = 0, 1, \ldots, n)$$

sowie P *das zugehörige Interpolationspolynom. Schließlich bezeichne für* $x \in [a,b]$

$$I(x) := [\min\{x_0, x\}, \ \max\{x_n, x\}] .$$

Wir behaupten: Zu jedem $x \in [a,b]$ *existiert ein* $\xi_x \in I(x)$, *so daß*

$$(7.1.21) \qquad f(x) - P(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi_x) \cdot p_{n+1}(x) .$$

Beweis. Für $x = x_k$ $(k \in \{0, 1, \ldots, n\})$ ist die Behauptung klar; es gilt nämlich dann einerseits $f(x_k) - P(x_k) = 0$ und andererseits $p_{n+1}(x_k) = 0$, mithin (7.1.21) mit beliebigem $\xi_x \in [a,b]$.

Im Fall $x \neq x_k$ $(k = 0, 1, \ldots, n)$ betrachten wir für $t \in [a,b]$

$$h(t) := f(t) - P(t) - (f(x) - P(x)) \frac{p_{n+1}(t)}{p_{n+1}(x)} .$$

Offenbar ist h eine $(n+1)$-mal differenzierbare Abbildung von $[a,b]$ in $\mathbb{R}$. Wir haben wegen

$$f(x_k) - P(x_k) = 0 , \quad p_{n+1}(x_k) = 0$$

zunächst

$$h(x_k) = 0 \quad (k = 0, 1, \ldots, n) ,$$

außerdem

$$h(x) = f(x) - P(x) - (f(x) - P(x)) \frac{p_{n+1}(x)}{p_{n+1}(x)} = 0 .$$

Zusammenfassend notieren wir:

$$\begin{cases} h \text{ ist } (n+1)\text{-mal differenzierbar in } [a,b], \\ h \text{ besitzt in } I(x) \text{ mindestens } (n+2) \text{ verschiedene Nullstellen.} \end{cases}$$

Nach dem Satz von Rolle liegt in jedem offenen Intervall, das durch zwei benachbarte Nullstellen von h begrenzt ist, mindestens eine Nullstelle von h'. Dies führt zu den Aussagen:

$$\begin{cases} h' \text{ ist n-mal differenzierbar in } [a,b], \\ h' \text{ besitzt in } I(x) \text{ mindestens } (n+1) \text{ verschiedene Nullstellen.} \end{cases}$$

Durch Induktion stellen wir fest:

$$\begin{cases} h^{(n)} \text{ ist in } [a,b] \text{ differenzierbar,} \\ h^{(n)} \text{ besitzt in } I(x) \text{ mindestens zwei verschiedene Nullstellen.} \end{cases}$$

Nach dem Satz von Rolle existiert also ein $\xi_x \in I(x)$ mit

$$h^{(n+1)}(\xi_x) = 0 .$$

Wegen Grad $(P) \leqq n$ ist die $(n+1)$-te Ableitung von P identisch Null; da $p_{n+1}(t)$ ein normiertes Polynom vom Grad $n+1$ ist, hat man

$$p_{n+1}^{(n+1)}(t) = (n+1)! .$$

Hieraus folgt

$$0 = h^{(n+1)}(\xi_x) = f^{(n+1)}(\xi_x) - (f(x) - P(x)) \frac{(n+1)!}{p_{n+1}(x)}$$

und durch einfaches Umformen schließlich die Gleichung (7.1.21).

Die Darstellung (7.1.21) läßt sich unmittelbar für eine Abschätzung des Restglieds verwenden.

(7.1.22) **Folgerung.** Ist

$$|f^{(n+1)}(t)| \leqq M \quad (t \in [a,b]) ,$$

so gilt für $x \in [a,b]$

$$(7.1.23) \qquad |f(x) - P(x)| \leqq \frac{M}{(n+1)!} \cdot |p_{n+1}(x)| .$$

Bei der praktischen Anwendung von (7.1.23) kann man auf die unter Umständen schwierige Berechnung von $f^{(n+1)}$ verzichten, sofern f in geeigneter Weise die Einschränkung einer holomorphen komplexwertigen Funktion ist. Hierzu notieren wir den

(7.1.24) **Satz.** *Es sei* f *eine holomorphe Funktion von einem Gebiet* $G \subset \mathbb{C}$ *in* $\mathbb{C}$, *dabei sei* $[a,b] \subset G$ *sowie für* $t \in [a,b]$ $f(t) \in \mathbb{R}$. *Ferner sei*

$$a \leqq x_0 < x_1 < \ldots < x_n \leqq b, \; f_i = f(x_i) \quad (i = 0, 1, \ldots, n) ,$$

P *das zugehörige Interpolationspolynom sowie*

$$I(x) = [\min \{x_0, x\}, \max \{x_n, x\}] .$$

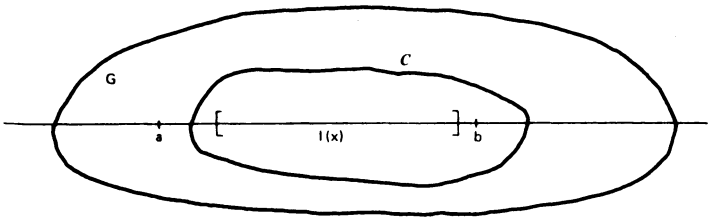
C sei ein einfach geschlossener Weg in G, $I(x)$ *liege im Inneren von* C, *und es bezeichne schließlich*

$$r(x) := \min\{|z-t| : t \in I(x), z \in (C)\},$$
$$M_C := \max\{|f(z)| : z \in (C)\},$$
$$l_C := \textit{die Länge von } C.$$

Unter diesen Voraussetzungen gilt die Abschätzung

$$|f(x) - P(x)| \leqq \frac{l_C \, M_C}{2\pi \cdot r(x)^{n+2}} |p_{n+1}(x)| . \tag{7.1.25}$$

Dem *Beweis* stellen wir zur Veranschaulichung der zahlreichen Bedingungen eine Skizze voran.



Nach der Cauchyschen Integralformel haben wir für $t \in I(x)$

$$f^{(n+1)}(t) = \frac{(n+1)!}{2\pi i} \int_C \frac{f(z)}{(z-t)^{n+2}} \, dz ,$$

folglich

$$|f^{(n+1)}(t)| \leqq \frac{(n+1)!}{2\pi} \cdot l_C \, \frac{1}{r(x)^{n+2}} M_C ,$$

mithin auf Grund der Abschätzung (7.1.23) die Behauptung.

Wir kehren zur Situation des Satzes (7.1.20) zurück; dabei betrachten wir speziell den Fall der linearen Interpolation, also $n = 1$. f sei in $[a,b]$ 2-mal stetig differenzierbar. Das Interpolationspolynom 1. Grades lautet – gemäß (7.1.5) bzw. (7.1.12) –

$$P_1(x) = f(x_1) \cdot \frac{x - x_0}{x_1 - x_0} + f(x_0) \cdot \frac{x - x_1}{x_0 - x_1} = f(x_0) + \Delta f[x_0, x_1] \cdot (x - x_0) .$$

Nach (7.1.23) gilt die Restgliedabschätzung

$$|f(x) - P_1(x)| \leqq \frac{|(x - x_0)(x - x_1)|}{2} \max_{t \in I(x)} |f''(t)| .$$

Im Fall $x_0 \leqq x \leqq x_1$ wird

$$|(x - x_0)(x - x_1)| = (x - x_0)(x_1 - x) .$$

Beachten wir, daß das geometrische Mittel durch das arithmetische Mittel majorisiert wird, so erhalten wir

$$\sqrt{(x - x_0)(x_1 - x)} \leqq \frac{(x - x_0) + (x_1 - x)}{2} = \frac{x_1 - x_0}{2}$$

und daher insgesamt die Abschätzung

$$(7.1.26) \qquad |f(x) - P_1(x)| \leqq \frac{(x_1 - x_0)^2}{8} \max_{t \in [x_0, x_1]} |f''(t)| \qquad (x \in [x_0, x_1]) .$$

Als *Beispiel* hierzu betrachten wir: es sei $f(x) = \sin x$ für

$$x = k \cdot h \qquad (k = 0, 1, 2, \ldots, N)$$

tabelliert und dabei die Schrittweite h mit

$$h = \frac{\pi}{2 \cdot N} \leqq 2 \cdot 10^{-2}$$

gewählt. Soll $\sin x$ für ein $x \in [0, \frac{\pi}{2}]$, $\neq k \cdot h$ mittels linearer Interpolation bestimmt werden, so wählt man zweckmäßigerweise x_0, x_1 als die benachbarten Stützstellen der Tafel, d.h. mit

$$x_0 < x < x_1 , \quad x_1 - x_0 = h$$

und bestimmt hierzu $P_1(x)$. Wegen

$$|f''(\xi)| = |-\sin \xi| \leqq 1$$

ergibt sich gemäß (7.1.26) die Fehlerabschätzung

$$|f(x) - P_1(x)| \leqq \frac{h^2}{8} \leqq \frac{1}{2} \cdot 10^{-4} ;$$

dies bedeutet, daß $f(x)$ durch $P_1(x)$ bis auf 4 Dezimalstellen nach dem Komma genau gegeben ist.

Entscheidend für die in (7.1.26) verwendete günstige Abschätzung von $|p_2(x)|$ ist die Tatsache, daß x zwischen x_0 und x_1 liegt. Auch bei größerem n erhält man ähnlich günstige Schranken für $|p_{n+1}(x)|$, sobald links und rechts von x etwa gleichviele Stützstellen liegen. Der Fall $n = 2$ wird in der Übungsaufgabe 7.4 diskutiert.

Bei der Interpolation von Funktionstafeln taucht das Problem auf, daß die tabellierten Funktionswerte gerundet sind und damit eine gestörte Interpolationsaufgabe gelöst wird. Demgemäß betrachten wir folgende Situation: vorgegeben seien $x_0, x_1, \ldots, x_n \in \mathbb{C}$, paarweise verschieden, ferner $f_0, f_1, \ldots, f_n$ sowie $\tilde{f}_0, \tilde{f}_1, \ldots, \tilde{f}_n \in \mathbb{C}$. Wir bezeichnen mit P bzw. $\tilde{P}$ die zugehörigen Interpolations-

polynome vom Grad $\leqq n$; nach (7.1.5) besitzt die Differenz die Darstellung

$$\tilde{P}(x) - P(x) = \sum_{k=0}^{n}{}' (\tilde{f}_k - f_k) \cdot l_k^n(x) .$$

Wir setzen

$$\epsilon := \max_{k=0}^{n} |\tilde{f}_k - f_k|$$

und folgern dann unmittelbar die Abschätzung

$$(7.1.27) \qquad |\tilde{P}(x) - P(x)| \leqq \epsilon \cdot \sum_{k=0}^{n}{}' |l_k^n(x)| \qquad (x \in \mathbb{C}) .$$

Im Fall der linearen Interpolation ($x_0 < x_1, \in \mathbb{R}$) ergibt sich für $x \in [x_0, x_1]$

$$|\tilde{P}_1(x) - P_1(x)| \leqq \epsilon \cdot \left(\left| \frac{x - x_0}{x_1 - x_0} \right| + \left| \frac{x - x_1}{x_0 - x_1} \right| \right) = \frac{\epsilon}{x_1 - x_0} (x - x_0 + x_1 - x) = \epsilon .$$

Bei sehr großem n ist jedoch damit zu rechnen, daß sich Fehler in den Eingabedaten unter erheblicher Verstärkung auf $|\tilde{P}(x) - P(x)|$ übertragen; zur Begründung zitieren wir den

(7.1.28) **Satz** *(G. Faber, S. N. Bernstein). Bezüglich reeller Stützstellen*

$$a \leqq x_0 < x_1 < \ldots < x_n \leqq b$$

gilt

$$\max_{x \in [a,b]} \sum_{k=0}^{n}{}' |l_k^n(x)| \geqq \frac{\ln(n+1)}{8 \cdot \sqrt{\pi}} .$$

Einen *Beweis* dieses Satzes findet man bei Natanson [41], Bd. III, S. 24.

Abschließend beschäftigt uns die Frage der gleichmäßigen Konvergenz der Interpolationspolynome bei wachsender Stützstellenzahl. Wir gehen aus von einem kompakten Intervall $[a,b] \subset \mathbb{R}$ und betrachten eine Folge von Stützstellen in $[a,b]$

$$(7.1.29) \qquad \left\{ \begin{array}{l} x_0^{(0)} , \\ x_0^{(1)} < x_1^{(1)} , \\ x_0^{(2)} < x_1^{(2)} < x_2^{(2)} , \\ x_0^{(3)} < x_1^{(3)} < x_2^{(3)} < x_3^{(3)} , \\ \quad \vdots \end{array} \right.$$

Zu einer vorgegebenen reellwertigen Funktion f auf $[a,b]$ bezeichnen wir mit $P_n(x)$ das Interpolationspolynom $\leqq$ n-ten Grades mit

$$P_n(x_k^{(n)}) = f(x_k^{(n)}) \qquad (k = 0, 1, \ldots, n) .$$

Daß für stetiges f die P_n im allgemeinen nicht gegen f konvergieren, besagt der

(7.1.30) **Satz** *(G. Faber). Zu jeder Stützstellenfolge* (7.1.29) *in* $[a,b]$ *existiert eine Funktion* $f \in C_0[a,b]$, *so daß die zugehörigen* $P_n(x)$ *nicht gleichmäßig gegen* f *konvergieren.*

Der *Beweis* dieses Satzes stützt sich im wesentlichen auf den oben zitierten Satz (7.1.28) und ist ebenfalls bei Natanson [41], Band III, S. 27 nachzulesen.

Es gilt jedoch der

(7.1.31) **Zusatz.** Ist f eine ganze Funktion, d.h. in ganz $\mathbb{C}$ holomorph, und auf $[a,b]$ reellwertig, so gilt bezüglich jeder Folge (7.1.29)

$$\max_{x \in [a,b]} |f(x) - P_n(x)| \to 0 \quad (n \to \infty) .$$

Beweis. Wir zeigen: für jedes $q > 0$ existiert ein $c \geqq 0$, so daß für alle $n \in \mathbb{N}$

$$\max_{x \in [a,b]} |f(x) - P_n(x)| \leqq c \cdot q^{n+1} \tag{7.1.32}$$

abschätzbar ist. Hieraus folgt dann die behauptete Konvergenz, da man insbesondere $q < 1$ vorgeben kann.

Zum Nachweis von (7.1.32) stützen wir uns auf den Satz (7.1.24), angewendet auf $G = \mathbb{C}$. Zu vorgegebenem $q > 0$ wählen wir die Kurve C so, daß bezüglich

$$r := \operatorname{dist}([a,b], (C)) = \min \{|t - z| : t \in [a,b], z \in (C)\}$$

die Ungleichung

$$\frac{b-a}{q} \leqq r$$

gilt. Dann haben wir nach (7.1.25) für alle $x \in [a,b]$

$$|f(x) - P_n(x)| \leqq \frac{l_C \, M_C}{2\pi} \, \frac{|p_{n+1}(x)|}{r^{n+2}} .$$

Hieraus folgt wegen

$$|p_{n+1}(x)| \leqq (b-a)^{n+1} \quad (x \in [a,b])$$

unmittelbar

$$|f(x) - P_n(x)| \leqq \frac{l_C \, M_C}{2\pi r} \left(\frac{b-a}{r}\right)^{n+1} \leqq \frac{l_C \, M_C}{2\pi r} \, q^{n+1} ,$$

also mit

$$c := \frac{l_C \, M_C}{2\pi r}$$

die Ungleichung (7.1.32).

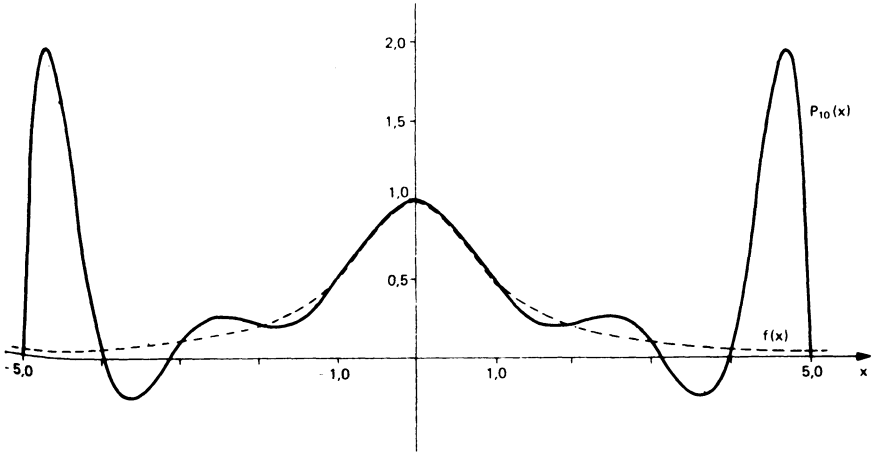
Nach dem angegebenen Beweis genügt es für die Konvergenz der P_n, wenn f in einem hinreichend großen Gebiet $G \subset \mathbb{C}$ holomorph ist. Daß diese Bedingung nicht weiter abzuschwächen ist, zeigt das *Beispiel*

$$(7.1.33)\quad \begin{cases} f(x) = \dfrac{1}{1+x^2}\,, \\ [a,b] = [-5,5]\,, \\ x_k^{(n)} = -5 + k \cdot \dfrac{10}{n} \quad (k = 0, 1, \ldots, n)\,. \end{cases}$$

In diesem Fall gilt nämlich (vgl. Isaacson/Keller [31], S. 287!)

$$\max_{x \in [-5,5]} |f(x) - P_n(x)| \to \infty \quad (n \to \infty)\,.$$

Die folgende Abbildung zeigt f(x) und das Interpolationspolynom $P_n(x)$ für $n = 10$.



7.2. Hermite-Interpolation

Als *Hermitesche Interpolationsaufgabe* bezeichnet man die folgende Problemstellung: Vorgegeben seien $x_0, x_1, \ldots, x_n \in \mathbb{C}$, paarweise verschieden, ferner $\alpha_0, \alpha_1, \ldots, \alpha_n \in \mathbb{N}$, $\neq 0$ und hierzu für $\rho \in \{0, 1, \ldots, \alpha_i - 1\}$, $i \in \{0, 1, \ldots, n\}$ Zahlen $f_i^{(\rho)} \in \mathbb{C}$. Gesucht ist ein Polynom H mit den Eigenschaften

$$(7.2.1)\quad \begin{cases} \mathrm{Grad}(H) \leqq m := (\alpha_0 + \alpha_1 + \ldots + \alpha_n) - 1\,, \\ H^{(\rho)}(x_i) = f_i^{(\rho)} \quad (\rho = 0, 1, \ldots, \alpha_i - 1;\ i = 0, 1, \ldots, n)\,. \end{cases}$$

Offensichtlich reduziert sich dieses Problem im Spezialfall $\alpha_i = 1$ $(i = 0, 1, \ldots, n)$ auf die in Abschnitt 7.1 ausführlich behandelte Newtonsche Interpolationsaufgabe. In Verallgemeinerung des Existenz- und Eindeutigkeitssatzes (7.1.2) notieren wir hier dementsprechend den

(7.2.2) **Satz.** *Das Hermitesche Interpolationsproblem* (7.2.1) *besitzt genau eine Lösung.*

Wir *beweisen* zunächst, daß es *höchstens* eine Lösung zu (7.2.1) gibt. Hierzu nehmen wir an, es seien $H, \tilde{H}$ Lösungen zu (7.2.1). Die Differenz $S = H - \tilde{H}$ ist dann ein Polynom vom Grad $\leqq m$ und besitzt in den x_i wegen

$$S^{(\rho)}(x_i) = H^{(\rho)}(x_i) - \tilde{H}^{(\rho)}(x_i) = f_i^{(\rho)} - f_i^{(\rho)} = 0 \quad (\rho = 0, 1, \ldots, \alpha_i - 1)$$

jeweils eine Nullstelle der Ordnung α_i, insgesamt also entsprechend der Ordnung mindestens $m + 1$ Nullstellen, woraus sofort $S = 0$, d.h. $H = \tilde{H}$ folgt.

Daß es *mindestens* eine Lösung zu (7.2.1) gibt, zeigen wir wiederum dadurch, daß wir eine derartige Lösung explizit angeben:

Dazu setzen wir zunächst

$$(7.2.3) \qquad w_\nu(x) = \prod_{\substack{\sigma = 0 \\ \sigma \neq \nu}}^{n} (x - x_\sigma)^{\alpha_\sigma} \qquad (\nu = 0, 1, \ldots, n)$$

sowie

$$(7.2.4) \qquad h_{\nu,\mu}(x) = w_\nu(x)\,(x - x_\nu)^\mu \quad (\mu = 0, 1, \ldots, \alpha_\nu - 1;\ \nu = 0, 1, \ldots, n)\,.$$

Hierzu stellen wir für $\rho = 0, 1, \ldots, \alpha_i - 1$; $i = 0, 1, \ldots, n$ die Beziehungen

$$(7.2.5) \qquad h_{\nu,\mu}^{(\rho)}(x_i) \begin{cases} \neq 0, & \text{falls } \nu = i,\ \mu = \rho\,, \\ = 0, & \text{falls } \nu \neq i \text{ oder } \nu = i,\ \mu > \rho\,, \end{cases}$$

fest; zur Begründung stützt man sich auf die Leibnizsche Regel zur Differentiation des Produktes zweier Funktionen – vgl. Forster [18], Bd. 1, S. 109 – und beachtet für $\nu = i$ die Definition (7.2.4) sowie für $i \neq \nu$ die Darstellung

$$h_{\nu,\mu}(x) = \Bigg[\prod_{\substack{\sigma = 0 \\ \sigma \neq \nu, i}}^{n} (x - x_\sigma)^{\alpha_\sigma}\,(x - x_\nu)^\mu\Bigg](x - x_i)^{\alpha_i} =: \tilde{h}(x)\,(x - x_i)^{\alpha_i}\,.$$

Mittels der $h_{\nu,\mu}$ definieren wir bei festem $\nu \in \{0, 1, \ldots, n\}$ rekursiv die Funktionen

$$(7.2.6) \qquad \begin{cases} l_{\nu,\alpha_\nu - 1}(x) := \dfrac{1}{h_{\nu,\alpha_\nu - 1}^{(\alpha_\nu - 1)}(x_\nu)}\, h_{\nu,\alpha_\nu - 1}(x)\,, \\ l_{\nu,\mu}(x) := \dfrac{1}{h_{\nu,\mu}^{(\mu)}(x_\nu)} \left\{ h_{\nu,\mu}(x) - \displaystyle\sum_{\sigma = \mu + 1}^{\alpha_\nu - 1} h_{\nu,\mu}^{(\sigma)}(x_\nu)\, l_{\nu,\sigma}(x) \right\} \\ (\mu = \alpha_\nu - 2, \alpha_\nu - 3, \ldots, 1, 0) \end{cases}$$

und zeigen

$$(7.2.7) \qquad l_{\nu,\mu}^{(\rho)}(x_i) = \delta_{i,\nu} \cdot \delta_{\rho,\mu} = \begin{cases} 1, & \text{falls } i = \nu, \rho = \mu, \\ 0, & \text{falls } i \neq \nu \text{ oder } \rho \neq \mu, \end{cases}$$

worin die Indizes die Bereiche

$$(7.2.8) \qquad \begin{cases} \mu = 0, 1, \ldots, \alpha_\nu - 1; \ \nu = 0, 1, \ldots, n; \\ \rho = 0, 1, \ldots, \alpha_i - 1; \ i = 0, 1, \ldots, n \end{cases}$$

durchlaufen. Den Beweis führen wir bei festem ν, i, ρ durch Induktion nach μ. Im Fall $\mu = \alpha_\nu - 1$ ist für $\nu = i$, $\rho = \mu\,(= \alpha_\nu - 1)$ nach (7.2.5)

$$l_{\nu,\alpha_\nu-1}^{(\alpha_\nu-1)}(x_\nu) = \frac{1}{h_{\nu,\alpha_\nu-1}^{(\alpha_\nu-1)}(x_\nu)} h_{\nu,\alpha_\nu-1}^{(\alpha_\nu-1)}(x_\nu) = 1 .$$

Für $\nu \neq i$ ergibt sich ebenfalls aus (7.2.5)

$$l_{\nu,\alpha_\nu-1}^{(\rho)}(x_i) = \frac{1}{h_{\nu,\alpha_\nu-1}^{(\alpha_\nu-1)}(x_\nu)} h_{\nu,\alpha_\nu-1}^{(\rho)}(x_i) = 0 ;$$

diese Beziehung ist natürlich auch richtig, falls $\nu = i$ sowie $\rho \neq \mu$, mithin gemäß (7.2.8) $\mu = \alpha_i - 1 > \rho$ gilt. Zum Schluß von $\mu + 1$ auf μ betrachten wir wie beim Induktionsanfang vorweg den Fall $\nu = i$, $\rho = \mu$. Hier hat man nach Induktionsannahme für $\sigma = \mu + 1, \mu + 2, \ldots, \alpha_\nu - 1$

$$l_{\nu,\sigma}^{(\mu)}(x_\nu) = 0 ,$$

was gemäß Definition (7.2.6) zu

$$l_{\nu,\mu}^{(\mu)}(x_\nu) = \frac{1}{h_{\nu,\mu}^{(\mu)}(x_\nu)} \{h_{\nu,\mu}^{(\mu)}(x_\nu) - 0\} = 1$$

führt. Ähnlich schließt man, falls $\nu \neq i$ oder $\nu = i$, $\rho < \mu$ gilt; denn dann ist einerseits auf Grund der Induktionsvoraussetzung für $\sigma = \mu + 1, \mu + 2, \ldots, \alpha_\nu - 1$

$$l_{\nu,\sigma}^{(\rho)}(x_i) = 0$$

sowie andererseits gemäß (7.2.5)

$$h_{\nu,\mu}^{(\rho)}(x_i) = 0 ,$$

also insgesamt

$$l_{\nu,\mu}^{(\rho)}(x_i) = \frac{1}{h_{\nu,\mu}^{(\mu)}(x_\nu)} \cdot \{0 - 0\} = 0 .$$

Etwas anders verläuft der Beweis im Fall $\nu = i$, $\rho > \mu$: Wegen

$$\rho \in \{\mu + 1, \mu + 2, \ldots, \alpha_\nu - 1\}$$

erhält man nach Induktionsannahme

$$\sum_{\sigma=\mu+1}^{\alpha_\nu-1} h_{\nu,\mu}^{(\sigma)}(x_\nu)\, l_{\nu,\sigma}^{(\rho)}(x_\nu) = h_{\nu,\mu}^{(\rho)}(x_\nu)$$

und infolgedessen

$$l_{\nu,\mu}^{(\rho)}(x_\nu) = \frac{1}{h_{\nu,\mu}^{(\mu)}(x_\nu)} \{h_{\nu,\mu}^{(\rho)}(x_\nu) - h_{\nu,\mu}^{(\rho)}(x_\nu)\} = 0\,;$$

damit ist (7.2.7) als richtig erkannt.

Zum Beweis von (7.2.2) bleibt

$$(7.2.9) \qquad H(x) := \sum_{\nu=0}^{n} \left\{ \sum_{\mu=0}^{\alpha_\nu-1} f_\nu^{(\mu)}\, l_{\nu,\mu}(x) \right\}$$

zu setzen und festzuhalten, daß die $l_{\nu,\mu}(x)$ Polynome höchstens m-ten Grades sind.

Im Spezialfall $n = 0$ liegt nur eine Interpolationsstelle vor; hier ergibt sich

$$(7.2.10) \qquad H(x) = \sum_{\mu=0}^{\alpha_0-1} \frac{1}{\mu!}(x - x_0)^\mu f_0^{(\mu)}\,.$$

Zu dieser Darstellung kommt man, wenn man der Reihe nach

$$(7.2.11) \qquad \begin{cases} w_0(x) = 1 \ \text{(leeres Produkt)}\,, \\ h_{0,\mu}(x) = (x - x_0)^\mu \quad (\mu = 0, 1, \ldots, \alpha_0 - 1)\,, \\ l_{0,\mu}(x) = \dfrac{1}{\mu!}(x - x_0)^\mu \quad (\mu = 0, 1, \ldots, \alpha_0 - 1) \end{cases}$$

berechnet, was als Übungsaufgabe 7.5, (i) empfohlen sei. Daß die auf der rechten Seite in (7.2.10) vorkommende Funktion im vorliegenden Fall das Hermitesche Problem (7.2.1) löst, erkennt man natürlich auch unmittelbar durch Differentiation.

Im Fall $\alpha_\nu = 1$ $(\nu = 0, 1, \ldots, n)$ reduziert sich die Hermitesche Interpolationsformel (7.2.9) auf die Lagrangesche Formel (7.1.5). Zum Beweis stellt man – vgl. Übungsaufgabe 7.5, (ii) – hier für $\nu = 0, 1, \ldots, n$ die Beziehungen

$$(7.2.12) \qquad \begin{cases} w_\nu(x) = \displaystyle\prod_{\substack{\sigma=0 \\ \sigma\neq\nu}}^{n} (x - x_\sigma)\,, \\ h_{\nu,0}(x) = w_\nu(x)\,, \\ l_{\nu,0}(x) = \dfrac{w_\nu(x)}{w_\nu(x_\nu)} = l_\nu^n(x) \end{cases}$$

fest; dabei ist natürlich von vornherein klar, daß hier das Lagrangesche Polynom (7.1.5) dem Problem (7.2.1) genügt.

Etwas eingehender diskutieren wir, und zwar im Hinblick auf die numerische Quadratur im Band 3, den Fall

(7.2.13) $\alpha_\nu = 2 \quad (\nu = 0, 1, \ldots, n)$.

Hier wird

$$w_\nu(x) = \prod_{\substack{\sigma = 0 \\ \sigma \neq \nu}}^{n} (x - x_\sigma)^2 ,$$

$$h_{\nu,0}(x) = w_\nu(x), \quad h_{\nu,1}(x) = w_\nu(x)\,(x - x_\nu) ,$$

insbesondere

$$h'_{\nu,0}(x_\nu) = w'_\nu(x_\nu) = \sum_{\substack{k = 0 \\ k \neq \nu}}^{n} 2(x_\nu - x_k) \prod_{\substack{\sigma = 0 \\ \sigma \neq k, \nu}}^{n} (x_\nu - x_\sigma)^2 = 2 w_\nu(x_\nu) \sum_{\substack{k = 0 \\ k \neq \nu}}^{n} \frac{1}{x_\nu - x_k}$$

sowie

$$h'_{\nu,1}(x_\nu) = w'_\nu(x_\nu)\,(x_\nu - x_\nu) + w_\nu(x_\nu) = w_\nu(x_\nu) ,$$

folglich nach (7.2.6)

$$l_{\nu,1}(x) = \frac{1}{w_\nu(x_\nu)} \cdot w_\nu(x)\,(x - x_\nu) ,$$

$$\begin{aligned} l_{\nu,0}(x) &= \frac{1}{w_\nu(x_\nu)} \cdot \{w_\nu(x) - h'_{\nu,0}(x_\nu)\, l_{\nu,1}(x)\} \\ &= \frac{1}{w_\nu(x_\nu)} \left\{ w_\nu(x) - 2 \cdot w_\nu(x_\nu) \sum_{\substack{k = 0 \\ k \neq \nu}}^{n} \frac{1}{x_\nu - x_k} \cdot \frac{1}{w_\nu(x_\nu)} w_\nu(x)\,(x - x_\nu) \right\} \\ &= \frac{w_\nu(x)}{w_\nu(x_\nu)} \left\{ 1 + 2 \cdot \sum_{\substack{k = 0 \\ k \neq \nu}}^{n} \frac{x_\nu - x}{x_\nu - x_k} \right\} . \end{aligned}$$

Eingesetzt in (7.2.9), führt dies hier zu der Interpolationsformel

$$(7.2.14) \qquad H(x) = \sum_{\nu = 0}^{n} \frac{w_\nu(x)}{w_\nu(x_\nu)} \left\{ f_\nu^{(0)} \left(1 + 2 \cdot \sum_{\substack{k = 0 \\ k \neq \nu}}^{n} \frac{x_\nu - x}{x_\nu - x_k} \right) + f_\nu^{(1)} (x - x_\nu) \right\} .$$

Zum Schluß kehren wir noch einmal zum allgemeinen Problem (7.2.1) zurück; dabei gehen wir davon aus, daß eine vorgegebene Funktion f interpoliert wird. Zur Darstellung und damit möglichen Abschätzung des hierbei entstehenden Fehlers notieren wir in Verallgemeinerung des Satzes (7.1.20) den

(7.2.15) **Satz.** *Es sei* f *eine* (m + 1)-mal *differenzierbare Abbildung von* $[a,b] \subset \mathbb{R}$ *in* $\mathbb{R}$. *Ferner sei (ohne Einschränkung)*

$$a \leqq x_0 < x_1 < \ldots < x_n \leqq b\,,$$
$$f_i^{(\rho)} = f^{(\rho)}(x_i) \quad (\rho = 0, 1, \ldots, \alpha_i - 1;\ i = 0, 1, \ldots, n)\,.$$

H *das zugehörige Hermitesche Interpolationspolynom* (7.2.5),

$$w(x) = \prod_{\sigma=0}^{n} (x - x_\sigma)^{\alpha_\sigma}\,,$$

schließlich für $x \in [a,b]$

$$I(x) := [\min\{x_0, x\},\ \max\{x_n, x\}]\,.$$

Wir behaupten: Zu jedem $x \in [a,b]$ *existiert ein* $\xi_x \in I(x)$, *so daß*

$$(7.2.16) \qquad f(x) - H(x) = \frac{1}{(m+1)!} f^{(m+1)}(\xi_x) \cdot w(x)\,.$$

Der *Beweis* verläuft völlig analog zu demjenigen des Satzes (7.1.20) und sei daher dem Leser (als Übungsaufgabe 7.6) überlassen.

7.3. Trigonometrische Interpolation

Wir bezeichnen mit $F_{2\pi}(\mathbb{R})$ die Menge der 2π-periodischen Funktionen von $\mathbb{R}$ in $\mathbb{C}$; ferner für $n \in \mathbb{N}$ mit

$$(7.3.1) \qquad T_n = \{g \in F_{2\pi}(\mathbb{R}):\ g(x) = \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos(kx) + b_k \sin(kx))\ (x \in \mathbb{R})\}\,,$$

d.h. die Gesamtheit der trigonometrischen Polynome vom Grad $\leqq n$.

Damit betrachten wir hier die folgende Aufgabe: zu $2n + 1$ verschiedenen Stützstellen

$$0 \leqq x_0 < x_1 < \ldots < x_{2n} < 2\pi$$

sowie $f_0, f_1, \ldots, f_{2n} \in \mathbb{C}$ ist eine Funktion g mit der Eigenschaft

$$(7.3.2) \qquad \begin{cases} g \in T_n \\ g(x_j) = f_j \quad (j = 0, 1, \ldots, 2n) \end{cases}$$

gesucht. Grundlegend ist der

(7.3.3) **Satz.** *Das trigonometrische Interpolationsproblem* (7.3.2) *ist eindeutig lösbar. Sind sämtliche* f_j *reell, so sind die Koeffizienten* a_k, b_k *der Lösung* $g \in T_n$ *ebenfalls reell.*

Zum *Beweis* gehen wir auf ein Problem der Polynom-Interpolation der in Abschnitt 7.1 betrachteten Art zurück.

Es bezeichne

$$K_1 := \{z \in \mathbb{C} : |z| = 1\}$$

sowie $F(K_1)$ die Gesamtheit aller Abbildungen von K_1 in $\mathbb{C}$.

(7.3.4) **Bemerkung.** Setzt man für $\tilde{f} \in F(K_1)$

$$\Phi(\tilde{f})(x) := \tilde{f}(e^{ix}) \quad (x \in \mathbb{R}),$$

so definiert Φ eine Bijektion von $F(K_1)$ auf $F_{2\pi}(\mathbb{R})$.

Beweis. Für $\tilde{f} \in F(K_1)$ ist wegen

$$e^{i(x+2\pi)} = e^{ix} \quad (x \in \mathbb{R})$$

$\Phi(\tilde{f})$ eine 2π-periodische Funktion; daher wird $F(K_1)$ durch Φ in $F_{2\pi}(\mathbb{R})$ abgebildet. Ist umgekehrt eine beliebige Funktion $f \in F_{2\pi}(\mathbb{R})$ vorgegeben, so wird auf Grund der Periodizität von f durch

$$(7.3.5) \qquad \tilde{f}(z) := f(x) \quad (z = e^{ix} \in K_1)$$

eindeutig eine Funktion $\tilde{f} \in F(K_1)$ definiert. Da (7.3.5) $\Phi(\tilde{f}) = f$ bedeutet, haben wir somit (7.3.4) bewiesen.

Geht man von g zu $\tilde{g} = \Phi^{-1}(g)$ und gleichzeitig von den x_j zu den – paarweise verschiedenen – $z_j = e^{ix_j}$ über, so gelangt man zu der zu (7.3.2) äquivalenten Interpolationsaufgabe: Gesucht ist eine Funktion $\tilde{g}$ mit der Eigenschaft

$$(7.3.6) \qquad \begin{cases} \tilde{g} \in \tilde{T}_n := \Phi^{-1}(T_n), \\ \tilde{g}(z_j) = f_j \quad (j = 0, 1, \ldots, 2n). \end{cases}$$

Hierzu zeigen wir zunächst den

(7.3.7) **Hilfssatz.** *Es gilt*

$$\tilde{T}_n = \{\tilde{g} \in F(K_1) : \tilde{g}(z) = \sum_{k=-n}^{+n} c_k z^k \quad (z \in K_1)\}.$$

Beweis. Es sei zunächst ein derartiges $\tilde{g} \in F(K_1)$ vorgegeben. Zur Bestimmung von $g = \Phi(\tilde{g})$ definieren wir $a_k, b_k \in \mathbb{C}$ durch

$$(7.3.8) \qquad \left.\begin{array}{l} c_0 = \dfrac{a_0}{2}, \\ c_k = \dfrac{1}{2}(a_k - i\,b_k), \\ c_{-k} = \dfrac{1}{2}(a_k + i\,b_k) \end{array}\right\} (k = 1, 2, \ldots, n).$$

Die dabei auftretenden (2,2)-Gleichungssysteme für die a_k, b_k sind offensichtlich eindeutig lösbar. Hiermit wird

$$g(x) = \frac{a_0}{2} + \sum_{k=1}^{n}{}' \frac{1}{2}(a_k - i\,b_k)e^{ikx} + \sum_{k=1}^{n}{}' \frac{1}{2}(a_k + i\,b_k)e^{-ikx}$$

$$= \frac{a_0}{2} + \sum_{k=1}^{n}{}' \left\{ a_k \cdot \frac{1}{2}(e^{ikx} + e^{-ikx}) + b_k \cdot \frac{1}{2i}(e^{ikx} - e^{-ikx}) \right\}$$

$$= \frac{a_0}{2} + \sum_{k=1}^{n}{}' (a_k \cos(kx) + b_k \sin(kx)) ;$$

dies bedeutet $g \in T_n$, also $\tilde{g} = \Phi^{-1}(g) \in \tilde{T}_n$.

Ist umgekehrt $\tilde{g} \in \tilde{T}_n$ vorgegeben, so ist $g = \Phi(\tilde{g}) \in T_n$. Daher besitzt g mit gewissen $a_k, b_k \in \mathbb{C}$ eine Darstellung wie in (7.3.1). Definiert man hierzu die c_k durch (7.3.8), so ist laut obiger Zwischenrechnung

$$g(x) = \sum_{k=-n}^{+n}{}' c_k e^{ikx}$$

und folglich, wie behauptet,

$$\tilde{g}(z) = \Phi^{-1}(g)(z) = \sum_{k=-n}^{+n}{}' c_k z^k .$$

Zu beachten bleibt schließlich der

(7.3.9) **Hilfssatz.** *Die (rationale) Interpolationsaufgabe* (7.3.6) *ist eindeutig lösbar. Sind sämtliche* f_j *reell, so gilt für die Koeffizienten* c_k *der Lösung* $\tilde{g}$

$$\overline{c_k} = c_{-k} \quad (k = 0, 1, \ldots, n) .$$

Beweis. Mittels der Substitution

$$P(z) = z^n \tilde{g}(z)$$

führt man (7.3.6) in das Problem

$$(*) \quad \begin{cases} P \text{ Polynom vom Grad} \leqq 2n , \\ P(z_j) = f_j \cdot z_j^n \quad (j = 0, 1, \ldots, 2n) \end{cases}$$

über. Gemäß (7.1.2) besitzt (*) genau eine Lösung; folglich existiert auch genau eine Lösung $\tilde{g}$ von (7.3.6).

Sind die $f_j \in \mathbb{R}$, so ergibt sich wegen $\overline{z_j} = z_j^{-1}$

$$f_j = \overline{f_j} = \overline{\tilde{g}(z_j)} = \sum_{k=-n}^{+n}{}' \overline{c_k} z_j^{-k} = \sum_{k=-n}^{+n}{}' \overline{c_{-k}} z_j^k ;$$

dementsprechend erfüllt die Funktion

$$\hat{g}(z) = \sum_{k=-n}^{+n} \overline{c_{-k}}\, z^k$$

ebenfalls die Bedingungen (7.3.6), woraus auf Grund der bereits bewiesenen Eindeutigkeit der Lösung

$$c_k = \overline{c_{-k}} \quad (k = 0, 1, \ldots, n)$$

folgt.

Für Anwendungen ist eine explizite Darstellung der Lösung g von (7.3.2) wünschenswert; in Analogie zur Lagrangeschen Interpolationsformel geben wir diese an im folgenden

(7.3.10) **Satz.** *Es bezeichne für* $k = 0, 1, \ldots, 2n$

$$(7.3.11) \qquad \omega_k(x) = \prod_{\substack{j=0 \\ j \neq k}}^{2n} \sin\left(\frac{x - x_j}{2}\right) .$$

Wir behaupten:

(i) *die* ω_k *sind trigonometrische Polynome vom Grad* $\leqq n$ *und besitzen die Eigenschaften*

$$(7.3.12) \qquad \omega_k(x_j) \begin{cases} = 0 & (j \neq k), \\ \neq 0 & (j = k). \end{cases}$$

(ii) *Die Funktion*

$$(7.3.13) \qquad g(x) = \sum_{k=0}^{n} f_k \cdot \frac{\omega_k(x)}{\omega_k(x_k)}$$

erfüllt die Interpolationsaufgabe (7.3.2).

Beweis. Die Aussage (ii) folgt trivialerweise aus (i). Zu (i) stellen wir zunächst fest, daß die ω_k jeweils Produkte von n Funktionen der Form

$$\begin{cases} \omega(x) = \sin\left(\frac{x-a}{2}\right) \cdot \sin\left(\frac{x-b}{2}\right), \\ a, b \in [0, 2\pi[\end{cases}$$

sind.

Da $\sin x$ in jedem halboffenen Intervall der Länge π genau eine Nullstelle besitzt, hat ω in $[0, 2\pi[$ genau die beiden Nullstellen a und b. Ferner gehört ω

zu T_1; zum Nachweis dieser Aussage stützt man sich auf die bekannten Additionstheoreme der trigonometrischen Funktionen und berechnet

$$\omega(x) = \frac{1}{2}\left\{\cos\left(\frac{a-b}{2}\right) - \cos\left(x - \frac{a+b}{2}\right)\right\}$$
$$= \frac{1}{2}\cos\left(\frac{a-b}{2}\right) - \frac{1}{2}\cos\left(\frac{a+b}{2}\right)\cdot\cos x - \frac{1}{2}\sin\left(\frac{a+b}{2}\right)\cdot\sin x\,.$$

Zu überlegen bleibt die Implikation

$$g_1 \in T_l,\ g_2 \in T_m \Rightarrow g := g_1 \cdot g_2 \in T_{l+m}\,.$$

Zu ihrem Beweis gehen wir zu den Funktionen

$$\tilde{g}_1 = \Phi^{-1}(g_1) \in \tilde{T}_l,\ \tilde{g}_1(z) = \sum_{\lambda=-l}^{+l}{}' c_\lambda^{(1)} z^\lambda\,,$$

$$\tilde{g}_2 = \Phi^{-1}(g_2) \in \tilde{T}_m,\ \tilde{g}_2(z) = \sum_{\mu=-m}^{+m}{}' c_\mu^{(2)} z^\mu$$

über. Für $\tilde{g} := \tilde{g}_1 \cdot \tilde{g}_2$ folgt die Darstellung

$$\tilde{g}(z) = \sum_{\kappa=-(m+l)}^{m+l}{}' c_\kappa z^\kappa\,,\quad c_\kappa = \sum_{\lambda+\mu=\kappa}{}' c_\lambda^{(1)} c_\mu^{(2)}\,,$$

mithin die Aussage $\tilde{g} \in \tilde{T}_{l+m}$ und daher wegen

$$g(x) = g_1(x)\,g_2(x) = \tilde{g}_1(e^{ix})\,\tilde{g}_2(e^{ix}) = \tilde{g}(e^{ix}) = \Phi(\tilde{g})(x)\,,$$

d.h. $g = \Phi(\tilde{g})$, wie behauptet, $g \in T_{l+m}$.

Die in diesem Abschnitt besprochene Interpolationsaufgabe bezieht sich stets auf eine ungerade Anzahl von Stützstellen. Eine entsprechende Aufgabe mit gerader Stützstellenzahl und etwa bezüglich Funktionen der Gestalt

$$(7.3.14)\qquad g(x) = \frac{a_0}{2} + \sum_{k=1}^{n-1}{}' (a_k \cos(kx) + b_k \sin(kx)) + a_n \cos(nx)$$

ist nicht für jede Wahl der Stützstellen in $[0, 2\pi[$ lösbar; dies bestätigt die Übungsaufgabe 7.7.

Für die numerische Anwendung besonders geeignet ist die trigonometrische Interpolation bei äquidistanten Stützstellen. Auf diese werden wir im Zusammenhang mit der Fourier-Analyse im Rahmen der Approximationstheorie in Band 3 eingehen.

7.4. Rationale Interpolation

Vorgegeben seien $x_0, x_1, \ldots, x_n \in \mathbb{C}$, und zwar paarweise verschieden, ferner $f_0, f_1, \ldots, f_n \in \mathbb{C}$ sowie $l, m \in \mathbb{N}$ mit $l + m = n$. Gesucht ist hier eine rationale Funktion

$$R(x) = \frac{P(x)}{Q(x)},$$

die den Bedingungen

$$(7.4.1) \quad \begin{cases} \text{Grad}(P) \leqq l, \text{Grad}(Q) \leqq m, Q \neq 0, \\ R(x_i) = f_i \quad (i = 0, 1, \ldots, n) \end{cases}$$

genügt. Anstelle dieser *rationalen Interpolationsaufgabe* betrachten wir zunächst das hiervon abgeleitete „linearisierte" Problem

$$(7.4.2) \quad \begin{cases} \text{Grad}(P) \leqq l, \ \text{Grad}(Q) \leqq m, Q \neq 0, \\ P(x_i) = f_i \cdot Q(x_i) \quad (i = 0, 1, \ldots, n) \end{cases}$$

und beweisen hierzu den

(7.4.3) **Satz.**

(i) *Die Aufgabe* (7.4.2) *ist lösbar.*

(ii) *Sind* P, Q *sowie* P_1, Q_1 *Lösungen zu* (7.4.2), *so folgt*

$$P Q_1 = P_1 Q .$$

Beweis.

(i) Indem wir

$$P(x) = \alpha_l x^l + \alpha_{l-1} x^{l-1} + \ldots + \alpha_0 ,$$
$$Q(x) = \beta_m x^m + \beta_{m-1} x^{m-1} + \ldots + \beta_0$$

ansetzen, erhalten wir nach (7.4.2) für die $n+2$ Koeffizienten $\alpha_0, \ldots, \alpha_l, \beta_0, \ldots, \beta_m$ ein homogenes lineares Gleichungssystem aus $n+1$ Gleichungen. Wie aus der linearen Algebra bekannt ist, besitzt ein System dieser Art mindestens eine nicht-triviale Lösung. Würden in einer derartigen Lösung sämtliche β_i verschwinden, so hätte man gemäß (7.4.2)

$$\text{Grad}(P) \leqq l \leqq n, \ P(x_i) = 0 \quad (i = 0, 1, \ldots, n) ;$$

im Widerspruch zu

$$(\alpha_0, \ldots, \alpha_l; \beta_0, \ldots, \beta_m) \neq (0, \ldots, 0; 0, \ldots, 0)$$

wäre dann auch $P = 0$.

(ii) Das Polynom $P Q_1 - P_1 Q$ hat einen Grad $\leqq n$ und genügt den Bedingungen

$$(P Q_1 - P_1 Q)(x_i) = f_i (Q(x_i) Q_1(x_i) - Q_1(x_i) Q(x_i)) = 0 \quad (i = 0, 1, \ldots, n)$$

infolgedessen ist es, wie zu zeigen war, das Nullpolynom.

Zur Eindeutigkeitsaussage (7.4.3), (ii) notieren wir ergänzend die folgende, allgemein gehaltene

(7.4.4) **Bemerkung.** Es seien P, Q sowie P_1, Q_1 jeweils teilerfremde Polynome über $\mathbb{C}$, $Q \neq 0$ und $Q_1 \neq 0$, ferner

$$(*) \quad P Q_1 = P_1 Q .$$

Dann existiert ein $\alpha \in \mathbb{C}$, $\neq 0$, so daß

$$P_1 = \alpha P, \quad Q_1 = \alpha Q .$$

Beweis. Aus (*) ergibt sich, daß Q das Produkt $P \cdot Q_1$ teilt. Da P, Q teilerfremd sind, muß Q – vgl. [16], S. 125 – auch Q_1 teilen. Analog erschließt man, daß Q_1 das Polynom Q teilen muß. Daraus folgt mit einen $\alpha \in \mathbb{C}$, $\neq 0$

$$Q_1 = \alpha Q .$$

Setzt man dies in (*) ein, so erhält man wegen $Q \neq 0$ auch

$$P_1 = \alpha P .$$

Bezüglich des ursprünglich gestellten Interpolationsproblems (7.4.1) erhalten wir den

(7.4.5) **Satz.**

(i) *Löst die rationale Funktion*

$$R = \frac{P}{Q}$$

die Aufgabe (7.4.1), *so erfüllen* P, Q *die Forderungen* (7.4.2).

(ii) *Erfüllen* P, Q *die Bedingungen* (7.4.2) *und sind sie zusätzlich teilerfremd, so löst*

$$R = \frac{P}{Q}$$

das Problem (7.4.1).

(iii) *Lösen die rationalen Funktionen*

$$\begin{cases} R = \dfrac{P}{Q}, & P, Q \text{ } \textit{teilerfremd}, \\ R_1 = \dfrac{P_1}{Q_1}, & P_1, Q_1 \text{ } \textit{teilerfremd} \end{cases}$$

beide das Problem (7.4.1), *so ist*

$$R = R_1 ;$$

in diesem Sinne ist (7.4.1) *– falls überhaupt – eindeutig lösbar.*

Beweis. Die Aussage (i) ist trivialerweise richtig; zu (ii) beachten wir, daß sich aus

$$Q(x_i) = 0, \quad P(x_i) = f_i \cdot Q(x_i)$$

die Beziehung $P(x_i) = 0$, also der Widerspruch

$x - x_i$ gemeinsamer Teiler von P und Q

ergäbe. Schließlich folgt (iii) aus (i), (7.4.3), (ii) sowie (7.4.4).

Die Voraussetzung in (ii), daß P, Q teilerfremd sind, ist notwendig. Andernfalls kann es nämlich vorkommen, daß P und Q in einem der vorgegebenen x_i eine gemeinsame Nullstelle besitzen, mithin R in x_i nicht definiert ist.

Haben P, Q beispielsweise den gemeinsamen Teiler $x - x_i$, so stellt sich die Frage, ob R in dem betreffenden x_i durch f_i stetig ergänzbar ist. Um dies zu untersuchen, betrachtet man

$$T = \mathrm{ggT}(P,Q) \quad (= \text{größter gemeinsamer Teiler von P und Q}),$$

$$\hat{P} = \frac{P}{T}, \quad \hat{Q} = \frac{Q}{T}$$

und hiermit als stetige Fortsetzung von R

$$\hat{R} = \frac{\hat{P}}{\hat{Q}}.$$

Dem Satz (7.4.5) entnimmt man unmittelbar als

(7.4.6) **Folgerung.** $\hat{R}$ löst genau dann die Aufgabe (7.4.1), wenn $\hat{P}$, $\hat{Q}$ die Eigenschaften (7.4.2) besitzen.

Ist $\hat{R}(x_i) \neq f_i$, d.h. hat $\hat{R}$ in x_i einen Pol oder nimmt dort einen komplexen Wert $\neq f_i$ an, so nennt man (x_i, f_i) einen *unerreichbaren Punkt* der Interpolationsaufgabe (7.4.1); die Aufgabe (7.4.1) ist dann natürlich *unlösbar.*

Interpolationsaufgaben, in denen unerreichbare Punkte auftreten, lassen sich leicht angeben. Wir betrachten folgende

(7.4.7) **Beispiele.** Wie man aus der Funktionentheorie (vgl. [54], S. 190) weiß, ist eine gebrochen-lineare Funktion

$$R(x) = \frac{\alpha_1 x + \alpha_0}{\beta_1 x + \beta_0}$$

entweder injektiv oder identisch einer Konstanten. Dies hat zur Folge, daß die Interpolationsaufgabe (7.4.1) mit

$$\begin{cases} l = m = 1, \; n = 2, \\ f_0 \neq f_1, \; f_1 = f_2 \end{cases}$$

keine Lösung besitzt. Als Zahlenbeispiel hierzu wählen wir

x_i	0	1	2
f_i	$\frac{1}{3}$	1	1

Dann erhalten wir aus (7.4.2) für die Koeffizienten von

$$P(x) = \alpha_0 + \alpha_1 x\,,\quad Q(x) = \beta_0 + \beta_1 x$$

das Gleichungssystem

$$\begin{cases} \alpha_0 & = \frac{1}{3}\beta_0 \\ \alpha_0 + \alpha_1 & = \beta_0 + \beta_1 \\ \alpha_0 + 2\alpha_1 & = \beta_0 + 2\beta_1\,. \end{cases}$$

Eine nicht-triviale Lösung dieses Systems ist

$$\alpha_0 = \beta_0 = 0\,,\quad \alpha_1 = \beta_1 = 1\,;$$

dies bedeutet

$$P(x) = Q(x) = x$$

sowie

$$\hat{R}(x) = 1\,;$$

mithin ist (x_0, f_0) unerreichbarer Punkt.

Die um ein geeignetes Paar (x_3, f_3) erweiterte Interpolationsaufgabe (mit $m = 2$, $l = 1$) ist jedoch lösbar; siehe Beispiel (7.4.23).

Ein Beispiel für einen unerreichbaren Punkt, in dem $R(x)$ einen Pol besitzt, bringt die Übungsaufgabe 7.8.

Obwohl Interpolationsaufgabe mit unerreichbaren Punkten leicht zu konstruieren sind, treten sie in der Praxis recht selten auf.

Zur praktischen Berechnung der rationalen Interpolierenden ist es im allgemeinen nicht zu empfehlen, die Koeffizienten von P und Q durch Lösen des homogenen Gleichungssystems in (7.4.2) zu bestimmen. Wir geben daher im Folgenden andere Berechnungsmethoden wieder; diese sind den zugkräftigen Algorithmen bei der Polynom-Interpolation nachgebildet.

Den Algorithmen von Neville und Aitken ist das nun zu entwickelnde, auf J. Stoer [51] zurückgehende Verfahren verwandt. Zu seiner Herleitung betrachten wir für $s, \mu, \nu \in \mathbb{N}$ mit

$$\mu \leqq l\,,\quad \nu \leqq m\,,\quad s + \mu + \nu \leqq n$$

die Interpolationsaufgaben

$$(I_s^{\mu,\nu}) \quad \begin{cases} R_s^{\mu,\nu} = \dfrac{P_s^{\mu,\nu}}{Q_s^{\mu,\nu}}\,,\ \mathrm{Grad}\,(P_s^{\mu,\nu}) \leqq \mu,\ \mathrm{Grad}\,(Q_s^{\mu,\nu}) \leqq \nu,\ Q_s^{\mu,\nu} \not\equiv 0\,, \\ R_s^{\mu,\nu}(x_i) = f_i \quad (i = s, s+1, \ldots, s+\mu+\nu) \end{cases}$$

bzw. die zugehörigen linearisierten Probleme

$$(L_s^{\mu,\nu}) \quad \begin{cases} \operatorname{Grad}(P_s^{\mu,\nu}) \leqq \mu, \ \operatorname{Grad}(Q_s^{\mu,\nu}) \leqq \nu, \ Q_s^{\mu,\nu} \neq 0 , \\ P_s^{\mu,\nu}(x_i) = f_i \cdot Q_s^{\mu,\nu}(x_i) \quad (i = s, s+1, \ldots, s+\mu+\nu) . \end{cases}$$

Dabei setzen wir voraus, daß die auftretenden $(I_s^{\mu,\nu})$ keine unerreichbaren Punkte besitzen; die dann im Sinne von (7.4.5), (iii) eindeutig existierende Lösung bezeichnen wir mit

$$\hat{R}_s^{\mu,\nu} = \frac{\hat{P}_s^{\mu,\nu}}{\hat{Q}_s^{\mu,\nu}} .$$

Weiter setzen wir

$$(7.4.8) \quad \begin{cases} \hat{P}_s^{\mu,\nu}(x) = \alpha_s^{\mu,\nu} x^{\mu} + \ldots , \\ \hat{Q}_s^{\mu,\nu}(x) = \beta_s^{\mu,\nu} x^{\nu} + \ldots \end{cases}$$

und notieren zunächst den

(7.4.9) **Hilfssatz.** *Es sei* $\mu + 1 \leqq l$, $\nu \leqq m$, $s + \mu + \nu + 1 \leqq n$.

(i) *Verschwindet eine der Zahlen* $\beta_{s+1}^{\mu,\nu}$, $\beta_s^{\mu,\nu}$, *so auch die andere, und es gilt*

$$\hat{R}_s^{\mu+1,\nu} = \hat{R}_{s+1}^{\mu,\nu} = \hat{R}_s^{\mu,\nu} .$$

(ii) *Sind die Zahlen* $\beta_{s+1}^{\mu,\nu}$, $\beta_s^{\mu,\nu}$ *beide von Null verschieden, so definieren wir*

$$(7.4.10) \quad \begin{cases} P_s^{\mu+1,\nu}(x) := \beta_s^{\mu,\nu}(x - x_s)\hat{P}_{s+1}^{\mu,\nu}(x) - \beta_{s+1}^{\mu,\nu}(x - x_{s+\mu+\nu+1})\hat{P}_s^{\mu,\nu}(x) , \\ Q_s^{\mu+1,\nu}(x) := \beta_s^{\mu,\nu}(x - x_s)\hat{Q}_{s+1}^{\mu,\nu}(x) - \beta_{s+1}^{\mu,\nu}(x - x_{s+\mu+\nu+1})\hat{Q}_s^{\mu,\nu}(x) . \end{cases}$$

Es existiert dann ein (von s, μ, ν *abhängiges)* $a \in \mathbb{C}$, $\neq 0$, *so daß*

$$P_s^{\mu+1,\nu} = a\hat{P}_s^{\mu+1,\nu}, \quad Q_s^{\mu+1,\nu} = a\hat{Q}_s^{\mu+1,\nu};$$

die $P_s^{\mu+1,\nu}$, $Q_s^{\mu+1,\nu}$ *sind also insbesondere teilerfremd und*

$$\hat{R}_s^{\mu+1,\nu} = \frac{P_s^{\mu+1,\nu}}{Q_s^{\mu+1,\nu}} .$$

Beweis.

(i) Ist $\beta_{s+1}^{\mu,\nu} = 0$, so sind die Polynome

$$\begin{cases} P_s^{\mu+1,\nu}(x) := (x - x_s)\hat{P}_{s+1}^{\mu,\nu}(x) , \\ Q_s^{\mu+1,\nu}(x) := (x - x_s)\hat{Q}_{s+1}^{\mu,\nu}(x) , \end{cases}$$

wie man leicht nachprüft, Lösungen von $(L_s^{\mu+1,\nu})$. Da $(I_s^{\mu+1,\nu})$ nach Voraussetzung keine unerreichbaren Punkte besitzt, erfüllen die reduzierten, d.h. teiler-

fremd gemachten Polynome

$$P_s^{\widehat{\mu+1,\nu}} = \hat{P}_{s+1}^{\mu,\nu}, \quad Q_s^{\widehat{\mu+1,\nu}} = \hat{Q}_{s+1}^{\mu,\nu}$$

ebenfalls die Bedingungen $(L_s^{\mu+1,\nu})$ und daher auch $(L_s^{\mu,\nu})$. Gemäß Satz (7.4.5), (ii), (iii) ist danach

$$\hat{R}_s^{\mu+1,\nu} = \hat{R}_{s+1}^{\mu,\nu} = \hat{R}_s^{\mu,\nu}.$$

Schließlich ergibt sich, z.B. auf Grund der Bemerkung (7.4.4), mit einem geeigneten $a \in \mathbb{C}, \neq 0$

$$\hat{P}_s^{\mu,\nu} = a \cdot \hat{P}_{s+1}^{\mu,\nu}, \quad \hat{Q}_s^{\mu,\nu} = a \cdot \hat{Q}_{s+1}^{\mu,\nu},$$

folglich insbesondere

$$\beta_s^{\mu,\nu} = a \cdot \beta_{s+1}^{\mu,\nu} = 0.$$

Entsprechend verfährt man im Fall $\beta_s^{\mu,\nu} = 0$.

(ii) Daraus, daß die Polynome $\hat{P}_{s+1}^{\mu,\nu}$, $\hat{Q}_{s+1}^{\mu,\nu}$ bzw. $\hat{P}_s^{\mu,\nu}$, $\hat{Q}_s^{\mu,\nu}$ den Beziehungen $(I_{s+1}^{\mu,\nu})$ und $(L_{s+1}^{\mu,\nu})$ bzw. $(I_s^{\mu,\nu})$ und $(L_s^{\mu,\nu})$ genügen, folgert man der Reihe nach:

Es gilt

$$\mathrm{Grad}(P_s^{\mu+1,\nu}) \leqq \mu + 1, \quad \mathrm{Grad}(Q_s^{\mu+1,\nu}) \leqq \nu,$$

letzteres, da der Koeffizient von $x^{\nu+1}$ in $Q_s^{\mu+1,\nu}$, nämlich

$$\beta_s^{\mu,\nu}\beta_{s+1}^{\mu,\nu} - \beta_{s+1}^{\mu,\nu}\beta_s^{\mu,\nu}$$

verschwindet. Weiter ist, wie man der Beziehung

$$Q_s^{\mu+1,\nu}(x_s) = -\beta_{s+1}^{\mu,\nu}(x_s - x_{s+\mu+\nu+1})\hat{Q}_s^{\mu,\nu}(x_s) \neq 0$$

entnimmt,

$$Q_s^{\mu+1,\nu} \neq 0.$$

Ferner wird, wie man leicht nachrechnet,

$$P_s^{\mu+1,\nu}(x_i) = f_i \cdot Q_s^{\mu+1,\nu}(x_i) \quad (i = s, s+1, \ldots, s+\mu+\nu+1).$$

Schließlich bleibt – man berücksichtige (7.4.6) sowie (7.4.5), (iii) – noch zu zeigen, daß die Polynome $P_s^{\mu+1,\nu}$, $Q_s^{\mu+1,\nu}$ teilerfremd sind. Hierzu nehmen wir das Gegenteil an, d.h. es existiere ein Polynom vom Grad $\geqq 1$, das sowohl $P_s^{\mu+1,\nu}$ als auch $Q_s^{\mu+1,\nu}$ teilt. Daraus würde folgen, daß die Grade der reduzierten Polynome die Ungleichungen

(*) $\mathrm{Grad}(P_s^{\widehat{\mu+1,\nu}}) \leqq \mu$,

(**) $\mathrm{Grad}(Q_s^{\widehat{\mu+1,\nu}}) \leqq \nu - 1$

erfüllten. Da $(I_s^{\mu+1,\nu})$ keine unerreichbaren Punkte hat, würden diese Polynome ebenfalls das Problem $(L_s^{\mu+1,\nu})$ lösen und infolgedessen nach (*), (**) natürlich auch der Aufgabe $(L_{s+1}^{\mu,\nu})$ genügen. Gemäß (7.4.4) ergäbe sich mit einem geeigneten $c \in \mathbb{C}, \neq 0$

$$\widehat{P_s^{\mu+1,\nu}} = c \cdot \hat{P}_{s+1}^{\mu,\nu}, \quad \widehat{Q_s^{\mu+1,\nu}} = c \cdot \hat{Q}_{s+1}^{\mu,\nu};$$

wegen (**) würde dies den Widerspruch $\beta_{s+1}^{\mu,\nu} = 0$ implizieren.

Um die im allgemeinen nicht explizit bekannten Größen $\beta_{s+1}^{\mu,\nu}$, $\beta_s^{\mu,\nu}$ aus (7.4.10) zu eliminieren, stützen wir uns auf den

(7.4.11) **Hilfssatz.** *Ist* $\nu \geqq 1$, *so gilt*

(i) $$\hat{R}_s^{\mu,\nu}(x) - \hat{R}_{s+1}^{\mu,\nu-1}(x) = -\alpha_{s+1}^{\mu,\nu-1}\beta_s^{\mu,\nu} \frac{(x-x_{s+1}) \cdot \ldots \cdot (x-x_{s+\mu+\nu})}{\hat{Q}_s^{\mu,\nu}(x)\,\hat{Q}_{s+1}^{\mu,\nu-1}(x)}$$

(ii) $$\hat{R}_{s+1}^{\mu,\nu}(x) - \hat{R}_{s+1}^{\mu,\nu-1}(x) = -\alpha_{s+1}^{\mu,\nu-1}\beta_{s+1}^{\mu,\nu} \frac{(x-x_{s+1}) \cdot \ldots \cdot (x-x_{s+\mu+\nu})}{\hat{Q}_{s+1}^{\mu,\nu}(x)\,\hat{Q}_{s+1}^{\mu,\nu-1}(x)}$$

Beweis. Wir haben

$$\hat{R}_s^{\mu,\nu}(x) - \hat{R}_{s+1}^{\mu,\nu-1}(x) = \frac{\hat{P}_s^{\mu,\nu}(x)\,\hat{Q}_{s+1}^{\mu,\nu-1}(x) - \hat{P}_{s+1}^{\mu,\nu-1}(x)\,\hat{Q}_s^{\mu,\nu}(x)}{\hat{Q}_s^{\mu,\nu}(x)\,\hat{Q}_{s+1}^{\mu,\nu-1}(x)} =: \frac{Z(x)}{N(x)},$$

worin das Zählerpolynom einen Grad $\leqq \mu + \nu$ und als Koeffizient von $x^{\mu+\nu}$ die komplexe Zahl

$$-\alpha_{s+1}^{\mu,\nu-1}\beta_s^{\mu,\nu}$$

besitzt. Ferner gilt für $i = s+1, \ldots, s+\mu+\nu$

$$Z(x_i) = f_i \cdot (\hat{Q}_s^{\mu,\nu}(x_i)\,\hat{Q}_{s+1}^{\mu,\nu-1}(x_i) - \hat{Q}_{s+1}^{\mu,\nu-1}(x_i)\,\hat{Q}_s^{\mu,\nu}(x_i)) = 0$$

und daher, wie behauptet

$$Z(x) = -\alpha_{s+1}^{\mu,\nu-1}\beta_s^{\mu,\nu}(x-x_{s+1}) \cdot \ldots \cdot (x-x_{s+\mu+\nu}).$$

Die Gleichung (ii) wird ebenso bewiesen.

Für den Übergang von $\hat{R}_s^{\mu,\nu}$, $\hat{R}_{s+1}^{\mu,\nu}$ auf $\hat{R}_s^{\mu,\nu+1}$ benötigt man die folgenden beiden Hilfssätze, die völlig analog zu (7.4.9) zu zeigen sind und deren Beweis wir daher dem Leser überlassen:

(7.4.9′) **Hilfssatz.** *Es sei* $\mu \leqq l$, $\nu + 1 \leqq m$, $s + \mu + \nu + 1 \leqq n$.

(i) *Verschwindet eine der Zahlen* $\alpha_{s+1}^{\mu,\nu}$, $\alpha_s^{\mu,\nu}$, *so auch die andere, und es gilt*

$$\hat{R}_s^{\mu,\nu+1} = \hat{R}_{s+1}^{\mu,\nu} = \hat{R}_s^{\mu,\nu}.$$

(ii) *Sind die Zahlen* $\alpha_{s+1}^{\mu,\nu}$, $\alpha_s^{\mu,\nu}$ *beide von Null verschieden, so definieren wir*

$$(7.4.10') \quad \begin{cases} P_s^{\mu,\nu+1}(x) := \alpha_s^{\mu,\nu}(x - x_s)\,\hat{P}_{s+1}^{\mu,\nu}(x) - \alpha_{s+1}^{\mu,\nu}(x - x_{s+\mu+\nu+1})\,\hat{P}_s^{\mu,\nu}(x), \\ Q_s^{\mu,\nu+1}(x) := \alpha_s^{\mu,\nu}(x - x_s)\hat{Q}_{s+1}^{\mu,\nu}(x) - \alpha_{s+1}^{\mu,\nu}(x - x_{s+\mu+\nu+1})\hat{Q}_s^{\mu,\nu}(x). \end{cases}$$

Es existiert dann ein (von s, μ, ν *abhängiges)* $a \in \mathbb{C}, \neq 0$, *so daß*

$$P_s^{\mu,\nu+1} = a \cdot \hat{P}_s^{\mu,\nu+1},\ Q_s^{\mu,\nu+1} = a \cdot \hat{Q}_s^{\mu,\nu+1};$$

die $P_s^{\mu,\nu+1}$, $Q_s^{\mu,\nu+1}$ *sind also insbesondere teilerfremd und*

$$\hat{R}_s^{\mu,\nu+1} = \frac{P_s^{\mu,\nu+1}}{Q_s^{\mu,\nu+1}}.$$

(7.4.11′) **Hilfssatz.** *Ist* $\mu \geqq 1$, *so gilt*

$$\text{(i)} \quad \hat{R}_s^{\mu,\nu}(x) - \hat{R}_{s+1}^{\mu-1,\nu}(x) = \beta_{s+1}^{\mu-1,\nu}\alpha_s^{\mu,\nu}\frac{(x - x_{s+1}) \cdot \ldots \cdot (x - x_{s+\mu+\nu})}{\hat{Q}_s^{\mu,\nu}(x)\,\hat{Q}_{s+1}^{\mu-1,\nu}(x)},$$

$$\text{(ii)} \quad \hat{R}_{s+1}^{\mu,\nu}(x) - \hat{R}_{s+1}^{\mu-1,\nu}(x) = \beta_{s+1}^{\mu-1,\nu}\alpha_{s+1}^{\mu,\nu}\frac{(x - x_{s+1}) \cdot \ldots \cdot (x - x_{s+\mu+\nu})}{\hat{Q}_{s+1}^{\mu,\nu}(x)\,\hat{Q}_{s+1}^{\mu-1,\nu}(x)}.$$

Nach diesen Vorbemerkungen beweisen wir in Analogie zu Satz (7.1.7), der Nevilleschen Formel, den

(7.4.12) **Satz.** (I) *Es sei* $\mu + 1 \leqq l$, $1 \leqq \nu \leqq m$, $s + \mu + \nu + 1 \leqq n$, *ferner* $\hat{x} \in \mathbb{C}, \neq x_0, x_1, \ldots, x_n$ *keine Polstelle von* $\hat{R}_{s+1}^{\mu,\nu}$, $\hat{R}_{s+1}^{\mu,\nu-1}$ *und* $\hat{R}_s^{\mu,\nu}$.

(i) *Falls dann eine der beiden Gleichungen*

$$\hat{R}_{s+1}^{\mu,\nu}(\hat{x}) = \hat{R}_{s+1}^{\mu,\nu-1}(\hat{x}),\ \hat{R}_s^{\mu,\nu}(\hat{x}) = \hat{R}_{s+1}^{\mu,\nu-1}(\hat{x})$$

zutrifft, so gilt auch die andere und außerdem

$$(7.4.13) \quad \hat{R}_s^{\mu+1,\nu}(\hat{x}) = \hat{R}_{s+1}^{\mu,\nu}(\hat{x}) = \hat{R}_s^{\mu,\nu}(\hat{x}).$$

(ii) *Andernfalls hat man*

$$(7.4.14) \quad \hat{R}_s^{\mu+1,\nu}(\hat{x}) = \hat{R}_{s+1}^{\mu,\nu}(\hat{x}) + \frac{\hat{R}_{s+1}^{\mu,\nu}(\hat{x}) - \hat{R}_s^{\mu,\nu}(\hat{x})}{\dfrac{\hat{x} - x_s}{\hat{x} - x_{s+\mu+\nu+1}}\left[1 - \dfrac{\hat{R}_{s+1}^{\mu,\nu}(\hat{x}) - \hat{R}_s^{\mu,\nu}(\hat{x})}{\hat{R}_{s+1}^{\mu,\nu}(\hat{x}) - \hat{R}_{s+1}^{\mu,\nu-1}(\hat{x})}\right] - 1}$$

(*eventuell* $= \infty$!).

(I′) *Es sei* $1 \leqq \mu \leqq l$, $\nu + 1 \leqq m$, $s + \mu + \nu + 1 \leqq n$, *ferner* $\hat{x} \in \mathbb{C}, \neq x_0, x_1, \ldots, x_n$ *keine Polstelle von* $\hat{R}_{s+1}^{\mu,\nu}$, $\hat{R}_{s+1}^{\mu-1,\nu}$ *und* $\hat{R}_s^{\mu,\nu}$.

(i′) *Falls dann eine der beiden Gleichungen*

$$\hat{R}_{s+1}^{\mu,\nu}(\hat{x}) = \hat{R}_{s+1}^{\mu-1,\nu}(\hat{x}),\ \hat{R}_s^{\mu,\nu}(\hat{x}) = \hat{R}_{s+1}^{\mu-1,\nu}(\hat{x})$$

zutrifft, so gilt auch die andere und außerdem

$$(7.4.13') \quad \hat{R}_s^{\mu,\nu+1}(\hat{x}) = \hat{R}_{s+1}^{\mu,\nu}(\hat{x}) = \hat{R}_s^{\mu,\nu}(\hat{x}) .$$

(ii') *Andernfalls hat man*

$$(7.4.14') \quad \hat{R}_s^{\mu,\nu+1}(\hat{x}) = \hat{R}_{s+1}^{\mu,\nu}(\hat{x}) + \frac{\hat{R}_{s+1}^{\mu,\nu}(\hat{x}) - \hat{R}_s^{\mu,\nu}(\hat{x})}{\dfrac{\hat{x}-x_s}{\hat{x}-x_{s+\mu+\nu+1}}\left[1 - \dfrac{\hat{R}_{s+1}^{\mu,\nu}(\hat{x}) - \hat{R}_s^{\mu,\nu}(\hat{x})}{\hat{R}_{s+1}^{\mu,\nu}(\hat{x}) - \hat{R}_{s+1}^{\mu-1,\nu}(\hat{x})}\right] - 1}$$

(*eventuell* $= \infty$!) .

Beweis.

(i) Es sei zunächst

$$\hat{R}_{s+1}^{\mu,\nu}(\hat{x}) = \hat{R}_{s+1}^{\mu,\nu-1}(\hat{x})$$

angenommen. Wegen (7.4.11), (ii) gilt dann

$$\alpha_{s+1}^{\mu,\nu-1} = 0 \quad \text{oder} \quad \beta_{s+1}^{\mu,\nu} = 0 .$$

Da im letzten Fall nach (7.4.9), (i) auch $\beta_s^{\mu,\nu}$ verschwindet, folgt in beiden Fällen mit (7.4.11)

$$\hat{R}_s^{\mu,\nu} = \hat{R}_{s+1}^{\mu,\nu} = \hat{R}_{s+1}^{\mu,\nu-1} .$$

$\hat{R}_s^{\mu,\nu}$ erfüllt also sowohl die Interpolationsaufgabe $(I_s^{\mu,\nu})$ als auch $(I_{s+1}^{\mu,\nu})$ und somit auch $(I_s^{\mu+1,\nu})$; dies impliziert auf Grund der eindeutigen Lösbarkeit

$$\hat{R}_s^{\mu+1,\nu} = \hat{R}_s^{\mu,\nu} .$$

Entsprechend wird die Aussage (i) im Fall

$$\hat{R}_s^{\mu,\nu}(\hat{x}) = \hat{R}_{s+1}^{\mu,\nu-1}(\hat{x})$$

sowie die Aussage (i') bewiesen.

(ii) Hier sind nach (i) und (7.4.11) die Zahlen $\alpha_{s+1}^{\mu,\nu-1}$, $\beta_{s+1}^{\mu,\nu}$, $\beta_s^{\mu,\nu}$ sämtlich von Null verschieden. Aus (7.4.10) folgt die Gleichung

$$\hat{R}_s^{\mu+1,\nu}(\hat{x}) = \frac{\beta_s^{\mu,\nu}(\hat{x}-x_s)\hat{P}_{s+1}^{\mu,\nu}(\hat{x}) - \beta_{s+1}^{\mu,\nu}(\hat{x}-x_{s+\mu+\nu+1})\hat{P}_s^{\mu,\nu}(\hat{x})}{\beta_s^{\mu,\nu}(\hat{x}-x_s)\hat{Q}_{s+1}^{\mu,\nu}(\hat{x}) - \beta_{s+1}^{\mu,\nu}(\hat{x}-x_{s+\mu+\nu+1})\hat{Q}_s^{\mu,\nu}(\hat{x})} .$$

Den auf der rechten Seite dieser Gleichung stehenden Bruch erweitern wir mit

$$-\alpha_{s+1}^{\mu,\nu-1} \cdot \frac{(\hat{x}-x_{s+1})\ldots(\hat{x}-x_{s+\mu+\nu})}{\hat{Q}_{s+1}^{\mu,\nu}(\hat{x})\,\hat{Q}_s^{\mu,\nu}(\hat{x})\,\hat{Q}_{s+1}^{\mu,\nu-1}(\hat{x})} \quad (\neq 0)$$

und erhalten nach (7.4.11)

$$(7.4.15)\quad \hat{R}_s^{\mu+1,\nu} = \frac{(\hat{x}-x_s)[\hat{R}_s^{\mu,\nu}-\hat{R}_{s+1}^{\mu,\nu-1}]\hat{R}_{s+1}^{\mu,\nu}-(\hat{x}-x_{s+\mu+\nu+1})[\hat{R}_{s+1}^{\mu,\nu}-\hat{R}_{s+1}^{\mu,\nu-1}]\hat{R}_s^{\mu,\nu}}{(\hat{x}-x_s)[\hat{R}_s^{\mu,\nu}-\hat{R}_{s+1}^{\mu,\nu-1}]-(\hat{x}-x_{s+\mu+\nu+1})[\hat{R}_{s+1}^{\mu,\nu}-\hat{R}_{s+1}^{\mu,\nu-1}]},$$

worin wir das Argument $\hat{x}$ hinter den Funktionsausdrücken zur Abkürzung weggelassen haben. Bezeichnet man den Nenner dieses Bruches mit N, so wird offenbar

$$\hat{R}_s^{\mu+1,\nu} = \frac{\{N+(\hat{x}-x_{s+\mu+\nu+1})[\hat{R}_{s+1}^{\mu,\nu}-\hat{R}_{s+1}^{\mu,\nu-1}]\}\hat{R}_{s+1}^{\mu,\nu}-(\hat{x}-x_{s+\mu+\nu+1})[\hat{R}_{s+1}^{\mu,\nu}-\hat{R}_{s+1}^{\mu,\nu-1}]\hat{R}_s^{\mu,\nu}}{N}$$

$$= \hat{R}_{s+1}^{\mu,\nu} + \frac{(\hat{x}-x_{s+\mu+\nu+1})[\hat{R}_{s+1}^{\mu,\nu}-\hat{R}_{s+1}^{\mu,\nu-1}][\hat{R}_{s+1}^{\mu,\nu}-\hat{R}_s^{\mu,\nu}]}{N}.$$

Es bleibt schließlich der Bruch durch

$$(\hat{x}-x_{s+\mu+\nu+1})[\hat{R}_{s+1}^{\mu,\nu}-\hat{R}_{s+1}^{\mu,\nu-1}] \quad (\neq 0)$$

zu kürzen, um die Gleichung (7.4.14) zu erhalten. Analog wird (7.4.14′) bewiesen.

Angemerkt sei, daß Zähler und Nenner der Brüche in den Formeln (7.4.14) bzw. (7.4.14′) nicht gleichzeitig Null werden, so daß diese Brüche mit Werten in $\mathbb{C} \cup \{\infty\}$ wohldefiniert sind.

Wichtig für die Anwendung des Satzes (7.4.12) ist die folgende

(7.4.16) **Ergänzung.** Setzt man zusätzlich für $x \in \mathbb{C}$

$$(7.4.17)\quad \hat{R}_{s+1}^{\mu,-1}(x) = \infty \quad (s = 0, \ldots, n-\mu-1;\ \mu = 0, \ldots, n-1)$$

sowie

$$(7.4.18)\quad \hat{R}_{s+1}^{-1,\nu}(x) = 0 \quad (s = 0, \ldots, n-\nu-1;\ \nu = 0, \ldots, n-1),$$

so bleiben im Fall $\nu = 0$ die Aussagen (i), (ii) des Satzes (7.4.12) und im Fall $\mu = 0$ die entsprechenden Aussagen (i′), (ii′) gültig.

Beweis. Nach Definition ist das Polynom $\hat{R}_s^{\mu,0}$ die Lösung der Interpolationsaufgabe (7.1.6). Daher sind gemäß (7.4.17) die beiden Gleichungen, d.h. die Prämisse der Aussage (7.4.12), (i) falsch, die Aussage (i) mithin wahr. Weiter erhält man für die $\hat{R}_s^{\mu,0}$ nach der Nevilleschen Formel in der Gestalt (7.1.9)

$$(7.4.19)\quad \hat{R}_s^{\mu+1,0}(\hat{x}) = \hat{R}_{s+1}^{\mu,0}(\hat{x}) + \frac{\hat{R}_{s+1}^{\mu,0}(\hat{x})-\hat{R}_s^{\mu,0}(\hat{x})}{\dfrac{\hat{x}-x_s}{\hat{x}-x_{s+\mu+1}}-1},$$

womit – berücksichtigt man noch (7.4.17) – die Behauptung in Bezug auf $\nu = 0$ bewiesen ist.

Zum Fall $\mu = 0$ überlegt man: nach Voraussetzung besitzen insbesondere die Interpolationsaufgaben $(I_s^{0,\nu})$ keine unerreichbaren Punkte; dies impliziert, daß die f_i entweder sämtlich verschwinden oder sämtlich von Null verschieden sind. Im ersten Fall ist

$$\hat{R}_s^{0,\nu+1} = \hat{R}_{s+1}^{0,\nu} = \hat{R}_s^{0,\nu} = 0\,,$$

folglich speziell die Aussage (i′) richtig. Im zweiten Fall gilt

$$\hat{R}_{s+1}^{0,\nu}(x) \neq 0\,,\quad \hat{R}_s^{0,\nu}(x) \neq 0 \quad (x \in \mathbb{C})\,;$$

daher ist hier die Gültigkeit der Formel (7.4.14′) nachzuweisen. Offensichtlich ist $\hat{R}_s^{0,\nu}(x)^{-1}$ die Lösung der Interpolationsaufgabe (7.1.6) mit ν statt μ und f_i^{-1} statt f_i. Demgemäß hat man wiederum nach der Nevilleschen Formel, diesmal in der ursprünglichen Gestalt (7.1.8),

$$\hat{R}_s^{0,\nu+1}(\hat{x})^{-1} = \frac{(\hat{x}-x_s)\,\hat{R}_{s+1}^{0,\nu}(\hat{x})^{-1} - (\hat{x}-x_{s+\nu+1})\,\hat{R}_s^{0,\nu}(\hat{x})^{-1}}{(\hat{x}-x_s)-(\hat{x}-x_{s+\nu+1})}\,,$$

folglich

$$\begin{aligned}\hat{R}_s^{0,\nu+1}(\hat{x}) &= \frac{(\hat{x}-x_s)-(\hat{x}-x_{s+\nu+1})}{\dfrac{\hat{x}-x_s}{\hat{R}_{s+1}^{0,\nu}(\hat{x})} - \dfrac{\hat{x}-x_{s+\nu+1}}{\hat{R}_s^{0,\nu}(\hat{x})}} \\ &= \hat{R}_{s+1}^{0,\nu}(\hat{x}) \cdot \frac{\hat{R}_s^{0,\nu}(\hat{x})\,(\hat{x}-x_s) - \hat{R}_s^{0,\nu}(\hat{x})\,(\hat{x}-x_{s+\nu+1})}{\hat{R}_s^{0,\nu}(\hat{x})\,(\hat{x}-x_s) - \hat{R}_{s+1}^{0,\nu}(\hat{x})\,(\hat{x}-x_{s+\nu+1})}\end{aligned}$$

und nach weiterer Zwischenrechnung, ähnlich derjenigen zur Herleitung der Formel (7.4.14) bzw. (7.4.14′), schließlich

$$(7.4.20)\qquad \hat{R}_s^{0,\nu+1}(\hat{x}) = \hat{R}_{s+1}^{0,\nu}(\hat{x}) + \frac{\hat{R}_{s+1}^{0,\nu}(\hat{x}) - \hat{R}_s^{0,\nu}(\hat{x})}{\dfrac{\hat{x}-x_s}{\hat{x}-x_{s+\nu+1}} \cdot \dfrac{\hat{R}_s^{0,\nu}(\hat{x})}{\hat{R}_{s+1}^{0,\nu}(\hat{x})} - 1}\,.$$

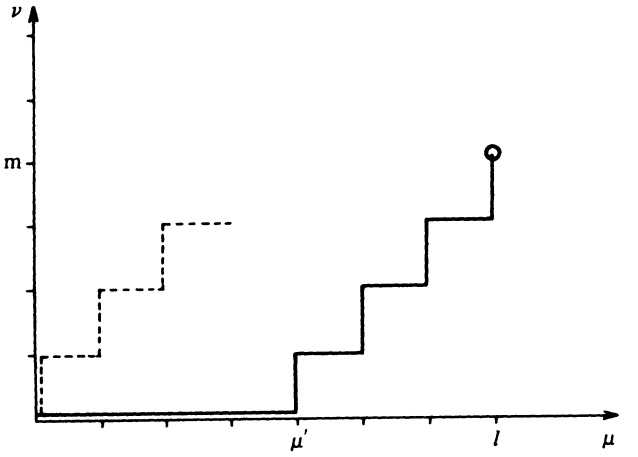
Wegen (7.4.18) stimmt diese Formel mit (7.4.14′) überein, was zu zeigen war.

Das *Stoersche Verfahren* besteht nun darin, die Lösung $\hat{R}$ der Interpolationsaufgabe (7.4.1) für ein festes $\hat{x} \neq x_0, x_1, \ldots, x_n$ dadurch zu berechnen, daß $\hat{R}(\hat{x}) = \hat{R}_0^{l,m}(\hat{x})$, von den Anfangswerten

$$\hat{R}_s^{0,0}(\hat{x}) = f_s \quad (s = 0, 1, \ldots, n)$$

ausgehend, mittels der Rekursionen in Satz (7.4.12) bestimmt wird. Dazu ist zu erklären, in welcher Reihenfolge die Rekursionen in (7.4.12), deren Anwendung eine Erhöhung des Index μ bzw. ν jeweils um 1 bewirkt, anzuwenden sind.

Ist beispielsweise $l > m$, so setzt man $\mu' = l - m + 1$ (oder auch $\mu' = l - m$) und berechnet zunächst $\hat{R}_s^{\mu',0}(\hat{x})$ für $s = 0, 1, \ldots, n - \mu'$ durch wiederholte Benutzung der Formeln (7.4.14), d.h. hier (7.4.18); danach erhöht man abwechselnd den Nenner- und Zählergrad um 1, indem man sich auf die Formeln (7.4.13), (7.4.14) bzw. (7.4.13'), (7.4.14') stützt. Wir erläutern das Vorgehen an Hand der nachstehenden Skizze:



Für die Extrapolationsverfahren, die im Zusammenhang mit der numerischen Quadratur und der Lösung gewöhnlicher Differentialgleichungen (Band 3) auftreten, haben sich Zähler- und Nennergrade l, m mit $l \leqq m \leqq l + 1$ als besonders geeignet erwiesen. Dementsprechend benutzt man die folgende Anordnung der (μ, ν)

(74.21) $\quad (0,0) \to (0,1) \to (1,1) \to (1,2) \to (2,2) \to \ldots \to (l, m)$,

die in obiger Skizze gestrichelt angedeutet ist.

Bei Vorliegen der Reihenfolge (7.4.21) läßt sich der Rechenvorgang in einfacher Weise beschreiben. Hierzu setzen wir

$$k := \mu + \nu,$$

$$\alpha_i := \hat{x} - x_i \quad (i = 0, 1, \ldots, n),$$

$$T_{i,k} := \hat{R}_i^{\mu,\nu}(\hat{x}) \quad (i = 0, \ldots, n - k;\ k = 0, 1, \ldots, n)$$

und erhalten dann gemäß Satz (7.4.12) in Verbindung mit der Ergänzung (7.4.16) den

(7.4.22) *Algorithmus.* Indem man, von

$$T_{i,-1} = 0, \quad T_{i,0} = f_i \quad (i = 0, 1, \ldots, n)$$

ausgehend, rekursiv für $k = 0, 1, \ldots, n-1$

$$T_{i,k+1} = \begin{cases} T_{i+1,k} & \text{im Fall } T_{i+1,k} = T_{i+1,k-1},\ T_{i,k} = T_{i+1,k-1}, \\ T_{i+1,k} + \dfrac{T_{i+1,k} - T_{i,k}}{\dfrac{\alpha_i}{\alpha_{i+k+1}}\left[1 - \dfrac{T_{i+1,k} - T_{i,k}}{T_{i+1,k} - T_{i+1,k-1}}\right] - 1} & \text{sonst} \end{cases}$$

$(i = 0, 1, \ldots, n-k-1)$

berechnet, erhält man – unter den oben genannten Voraussetzungen – nach n Schritten

$$\hat{R}(\hat{x}) = \hat{R}_0^{l,m}(\hat{x}) = T_{0,n}.$$

Zur Darstellung verwenden wir das Schema

$$\begin{array}{cccccc} 0 & f_0 & & & & \\ & & T_{0,1} & & & \\ 0 & f_1 & & T_{0,2} & & \\ & & T_{1,1} & & T_{0,3} & \\ 0 & f_2 & & T_{1,2} & & \\ & & T_{2,1} & & & \\ \vdots & f_3 & \vdots & \vdots & \vdots & \ddots \\ \vdots & \vdots & \vdots & \vdots & & \\ \vdots & & & & & \end{array};$$

hierin stehen jeweils die Ecken einer Raute in Beziehung zueinander.

Wir weisen noch einmal ausdrücklich darauf hin, daß nach Satz (7.4.12) von den in (7.4.22) erwähnten Gleichungen

$$T_{i+1,k} = T_{i+1,k-1},\ T_{i,k} = T_{i+1,k-1}$$

entweder keine oder beide erfüllt sind. Sollte bei Durchführung des Algorithmus beispielsweise der Fall

$$T_{i+1,k} = T_{i+1,k-1},\ T_{i,k} \neq T_{i+1,k-1}$$

auftreten, so bedeutet dies, daß eine der Interpolationsaufgaben $(I_s^{\mu,\nu})$ keine Lösung besitzt und daher die Anwendung des Algorithmus nicht gerechtfertigt ist.

(7.4.23) **Zahlenbeispiel.** Wir wollen $\hat{R}_0^{1,2}(\hat{x})$ an der Stelle $\hat{x} = 4$ berechnen, wobei $\hat{R}_0^{1,2}$ durch die Interpolationsvorschrift

x_i	0	1	2	3
f_i	$\frac{1}{3}$	1	1	$\frac{2}{3}$

bestimmt sei. Bei Verwendung des Algorithmus (7.4.22) erhalten wir gemäß obigem Schema die Werte

$$\begin{array}{ccccc} T_{i,-1} & T_{i,0} & T_{i,1} & T_{i,2} & T_{i,3} \\ & \frac{1}{3} & & & \\ 0 & & -\frac{1}{5} & & \\ & 1 & & 1 & \\ 0 & & 1 & & 1 \\ & 1 & & 1 & \\ 0 & & \frac{1}{2} & & \\ & \frac{2}{3} & & & \end{array}$$

Hier haben wir

$$T_{1,1} = T_{1,0}, \quad \text{jedoch} \quad T_{0,1} \neq T_{1,0}$$

und außerdem

$$T_{1,1} = T_{2,0}, \quad \text{jedoch} \quad T_{2,1} \neq T_{2,0}.$$

Wie auch im Beispiel (7.4.7) schon erwähnt, sind die Interpolationsaufgaben $(I_0^{1,1})$ sowie $(I_1^{1,1})$ nicht lösbar. Wenn wir dennoch die $T_{i,2}$ und $T_{0,3}$ nach den Formeln (7.4.22) – in einer Form (7.4.15) – berechnen, so erhalten wir als Lösung

$$\hat{R}_0^{1,2}(4) = T_{0,3} = 1.$$

Dieser Wert ist jedoch falsch: man überzeugt sich nämlich, daß die gestellte Interpolationsaufgabe durch die Funktion

$$\hat{R}_0^{1,2}(x) = \frac{x+1}{x^2 - 2x + 3}$$

gelöst wird, so daß

$$\hat{R}_0^{1,2}(4) = \frac{5}{11}$$

wird. – Bei Vertauschung der Stützstellen liefert das Stoersche Verfahren den richtigen Wert, vgl. Übungsaufgabe 7.9.

Ähnlich wie bei der Polynom-Interpolations die Algorithmen von Neville und Aitken, so ist hier das Stoersche Verfahren weniger brauchbar, wenn mehrere Funktionswerte der rationalen Interpolierenden berechnet werden müssen. Wir entwickeln dafür im Folgenden ein anderes Berechnungsverfahren, das, wie wir sehen werden, mit der Newtonschen Methode (7.1.16) verwandt ist. Dieses liefert hier die Lösung von (7.4.2) bzw. (7.4.1) mittels einfacher linearer Rekursionen bzw. als endlichen Kettenbruch *(Thieleschen Kettenbruch)*.

Hierzu führen wir zunächst die *inversen Differenzen* rekursiv durch

$$(7.4.24)\quad \begin{cases} \varphi_0(x_i) = f_i \quad (0 \leqq i \leqq n)\,, \\ \varphi_k(x_0, \ldots, x_{k-1}, x_i) = \dfrac{x_i - x_{k-1}}{\varphi_{k-1}(x_0, \ldots, x_{k-2}, x_i) - \varphi_{k-1}(x_0, \ldots, x_{k-1})} \\ (1 \leqq k \leqq i \leqq n) \end{cases}$$

ein; zu ihrer Berechnung benutzt man das

(7.4.25) **Schema.**

$$\begin{array}{llllll} x_0 & \varphi_0(x_0) & & & & \\ x_1 & \varphi_0(x_1) & \varphi_1(x_0, x_1) & & & \\ x_2 & \varphi_0(x_2) & \varphi_1(x_0, x_2) & \varphi_2(x_0, x_1, x_2) & & \\ x_3 & \varphi_0(x_3) & \varphi_1(x_0, x_3) & \varphi_2(x_0, x_1, x_3) & \varphi_3(x_0, x_1, x_2, x_3) & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{array}$$

Dabei setzen wir hier und im weiteren voraus, daß die Nenner in (7.4.24) nicht verschwinden, d.h. daß die φ_i in (7.4.24) als komplexe Zahlen ($\neq 0$) definiert sind. Wie das Beispiel (7.4.47) zeigt, schließt diese Voraussetzung nicht aus, daß unerreichbare Punkte auftreten. Auf die Situation, daß in (7.4.24) durch Null dividiert wird, so daß mit dem Wert ∞ zu rechnen ist, gehen wir in Übungsaufgabe 7.10 gesondert ein.

Zur Abkürzung setzen wir

$$(7.4.26)\quad \alpha_k := \varphi_k(x_0, x_1, \ldots, x_k) \quad (0 \leqq k \leqq n)$$

sowie für $x \in \mathbb{C}$

$$(7.4.27)\quad M_{k,i}(x) := \begin{pmatrix} \alpha_k & x - x_k \\ 1 & 0 \end{pmatrix} \cdot \ldots \cdot \begin{pmatrix} \alpha_{i-1} & x - x_{i-1} \\ 1 & 0 \end{pmatrix} \quad (0 \leqq k \leqq i \leqq n)\,;$$

dabei bedeute speziell $M_{i,i}(x)$ die (2,2)-Einheitsmatrix I_2. Schließlich definieren wir, ebenfalls für $x \in \mathbb{C}$

$$(7.4.28)\quad \begin{pmatrix} Z_{k,i}(x) \\ N_{k,i}(x) \end{pmatrix} := M_{k,i}(x) \begin{pmatrix} \alpha_i \\ 1 \end{pmatrix} \quad (0 \leqq k \leqq i \leqq n)\,.$$

Wir formulieren hierzu den

(7.4.29) **Hilfssatz.**

(i) *Bei festem* $i \in \{0, 1, \ldots, n\}$ *genügt die Folge der*

$$\begin{pmatrix} Z_{k,i}(x) \\ N_{k,i}(x) \end{pmatrix} \quad (k = i, i-1, \ldots, 1)$$

der homogenen linearen Differenzengleichung erster Ordnung

$$(7.4.30)\quad \begin{cases} \begin{pmatrix} Z_{k-1,i}(x) \\ N_{k-1,i}(x) \end{pmatrix} = \begin{pmatrix} \alpha_{k-1} & x - x_{k-1} \\ 1 & 0 \end{pmatrix} \begin{pmatrix} Z_{k,i}(x) \\ N_{k,i}(x) \end{pmatrix} & (k = i, i-1, \ldots, 1) \\ \begin{pmatrix} Z_{i,i}(x) \\ N_{i,i}(x) \end{pmatrix} = \begin{pmatrix} \alpha_i \\ 1 \end{pmatrix} \end{cases}$$

und ist durch diese bestimmt.

(ii) *Für* $i = 0, 1, \ldots, n$ *bezeichne*

$$D_i := \{x \in \mathbb{C} : \forall\, k = 0, 1, \ldots, i \quad N_{k,i}(x) \neq 0\}$$

und weiter für $x \in D_i$

$$S_{k,i}(x) := \frac{Z_{k,i}(x)}{N_{k,i}(x)} \quad (0 \leqq k \leqq i)\,.$$

Es ist dann für $x \in D_i$

$$S_{k,i}(x) \neq 0 \quad (1 \leqq k \leqq i)\,;$$

ferner gilt die nicht-lineare Rekursion

$$(7.4.31)\quad S_{k-1,i}(x) = \alpha_{k-1} + \frac{x - x_{k-1}}{S_{k,i}(x)} \quad (k = i, i-1, \ldots, 1)\,.$$

(iii) *Für* $i = 0, 1, \ldots, n$ *hat man* $x_i \in D_i$ *und*

$$(7.4.32)\quad S_{k,i}(x_i) = \varphi_k(x_0, \ldots, x_{k-1}, x_i) \quad (0 \leqq k \leqq i)\,,$$

mithin insbesondere

$$(7.4.33)\quad S_{0,i}(x_i) = \varphi_0(x_i) = f_i\,.$$

Zum *Beweis* von (i) beachten wir die Beziehung

$$(7.4.34)\quad M_{k,i}(x) = M_{k,j}(x)\, M_{j,i}(x) \quad (k \leqq j \leqq i)\,.$$

Um die Aussage (ii), d.h. die Rekursion (7.4.31) zu begründen, schreibt man sich die Differenzengleichung (7.4.30) komponentenweise, also in der Form

$$(7.4.35)\quad \begin{cases} Z_{k-1,i}(x) = \alpha_{k-1} Z_{k,i}(x) + (x - x_{k-1})\, N_{k,i}(x)\,, \\ N_{k-1,i}(x) = Z_{k,i}(x) \end{cases}$$

auf. Es bleibt dann die erste dieser Gleichungen durch $Z_{k,i}(x)$ zu dividieren und dabei die zweite zu beachten. Angemerkt sei noch, daß für jedes $x \in \mathbb{C}$

$$(7.4.36)\quad N_{i,i}(x) = 1,\ N_{i-1,i}(x) = Z_{i,i}(x) = \alpha_i \ (\neq 0)$$

und demgemäß

$$(7.4.37)\quad D_i = \{x \in \mathbb{C} : \forall\, k = 0, 1, \ldots, i-2 \quad N_{k,i}(x) \neq 0\}$$

ist.

Die Behauptung (iii) beweisen wir durch Induktion nach k: Für $k = i$ (Induktionsanfang) hat man nach (7.4.36) $N_{i,i}(x_i) \neq 0$ sowie

$$S_{i,i}(x_i) = \frac{Z_{i,i}(x_i)}{N_{i,i}(x_i)} = \alpha_i = \varphi_i(x_0, x_1, \ldots, x_i) .$$

Zum Schluß von k auf $k-1$ stellen wir zunächst fest, daß gemäß der Definition (7.4.24) für $1 \leqq k \leqq i$

$$\varphi_k(x_0, \ldots, x_{k-1}, x_i) \neq 0$$

und weiter

$$(7.4.38) \qquad \varphi_{k-1}(x_0, \ldots, x_{k-2}, x_i) = \varphi_{k-1}(x_0, \ldots, x_{k-1}) + \frac{x_i - x_{k-1}}{\varphi_k(x_0, \ldots, x_{k-1}, x_i)}$$

gilt. Aus der Induktionsannahme folgert man dann auf Grund der Rekursion (7.4.35) der Reihe nach

$$N_{k-1,i}(x_i) = Z_{k,i}(x_i) = \varphi_k(x_0, \ldots, x_{k-1}, x_i)\, N_{k,i}(x_i) \neq 0$$

sowie

$$\begin{aligned} Z_{k-1,i}(x_i) &= \alpha_{k-1} Z_{k,i}(x_i) + (x_i - x_{k-1})\, N_{k,i}(x_i) \\ &= \{\varphi_{k-1}(x_0, \ldots, x_{k-1})\, \varphi_k(x_0, \ldots, x_{k-1}, x_i) + (x_i - x_{k-1})\}\, N_{k,i}(x_i) , \end{aligned}$$

insgesamt also

$$S_{k-1,i}(x_i) = \frac{Z_{k-1,i}(x_i)}{N_{k-1,i}(x_i)} = \varphi_{k-1}(x_0, \ldots, x_{k-1}) + \frac{x_i - x_{k-1}}{\varphi_k(x_0, \ldots, x_{k-1}, x_i)}$$

womit nach (7.4.38) die Beziehung (7.4.32) für $k-1$, d.h. die Induktionsbehauptung bewiesen ist.

In erster Linie interessiert uns im Folgenden

$$(7.4.39) \qquad \begin{cases} Z_k := Z_{k,n},\ N_k := N_{k,n},\ S_k := S_{k,n} & (0 \leqq k \leqq n) , \\ P := Z_0,\ Q := N_0,\ R := \dfrac{P}{Q} = S_0 . \end{cases}$$

Hierzu notieren wir den

(7.4.40) **Satz.**

(i) P, Q *sind Polynome der Gestalt*

$$P(x) = \sum_{\mu=0}^{l} a_\mu x^\mu , \quad Q(x) = \sum_{\nu=0}^{m} b_\nu x^\nu$$

mit

$$l = n - m,\ m = \left[\frac{n}{2}\right] ;$$

dabei ist im Fall n *gerade,* $= 2m$

$$a_m = \sum_{i=0}^{m}{}' \varphi_{2i}(x_0, \ldots, x_{2i}), \quad b_m = 1$$

sowie im Fall n *ungerade,* $= 2m + 1$

$$a_{m+1} = 1, \quad b_m = \sum_{i=0}^{m}{}' \varphi_{2i+1}(x_0, \ldots, x_{2i+1}).$$

(ii) P, Q *lösen bezüglich der in* (i) *angegebenen* l, m *das linearisierte Interpolationsproblem* (7.4.2).

(iii) *Es ist* $Q(x_n) \neq 0$, *mithin*

$$R(x_n) = f_n .$$

Gilt auch für $i = 0, 1, \ldots, n-1$ $Q(x_i) \neq 0$, *so genügt* R *bezüglich der in* (i) *angegebenen* l, m *dem rationalen Interpolationsproblem* (7.4.1).

Beweis.

(i) Nach (7.4.30) bzw. (7.4.35) sind die Z_k, N_k die Lösungen der Rekursionen

$$(7.4.41) \quad \begin{cases} Z_{k-1}(x) = \alpha_{k-1} Z_k(x) + (x - x_{k-1}) N_k(x), \\ N_{k-1}(x) = Z_k(x) \end{cases} \qquad (k = n, n-1, \ldots, 1)$$

zu den Anfangsbedingungen

$$(7.4.42) \quad Z_n(x) = \alpha_n, \; N_n(x) = 1 .$$

Im Falle $n \geqq 2$ ergeben sich aus (7.4.41) durch Zusammenfassung der Gleichungen für k und $k-1$ die Beziehungen

$$(7.4.43) \quad \begin{cases} Z_{k-2}(x) = (\alpha_{k-1}\alpha_{k-2} + (x - x_{k-2})) Z_k(x) + \alpha_{k-2}(x - x_{k-1}) N_k(x), \\ N_{k-2}(x) = \alpha_{k-1} Z_k(x) + (x - x_{k-1}) N_k(x) \\ (k = n, n-1, \ldots, 2) . \end{cases}$$

Weiter sei zunächst $n = 2m$. Im Fall $m = 0$ ist die Behauptung unmittelbar klar. Für $m > 0$ folgt aus (7.4.43) wegen

$$\mathrm{Grad}(Z_n) = \mathrm{Grad}(N_n) = 0$$

für $k = 0, 1, \ldots, m$ induktiv

$$\mathrm{Grad}(Z_{2(m-k)}) \leqq k, \quad \mathrm{Grad}(N_{2(m-k)}) = k .$$

Genauer erschließen wir auf diese Weise

$$Z_{2(m-k)}(x) = \Big(\sum_{i=m-k}^{m}{}' \alpha_{2i} \Big) x^k + \ldots, \quad N_{2(m-k)}(x) = x^k + \ldots .$$

Setzen wir hierin $k = m$, so ergibt sich die Behauptung (i) für $n = 2m$.

Ist n ungerade, $= 2m + 1$, so leiten wir aus (7.4.43) ähnlich wie oben für $k = 0, 1, \ldots, m$ die Beziehungen

$$\mathrm{Grad}(Z_{2(m-k)+1}) \leqq k, \quad \mathrm{Grad}(N_{2(m-k)+1}) = k,$$

$$Z_{2(m-k)+1}(x) = \Big(\sum_{i=m-k}^{m} \alpha_{2i+1} \Big) x^k + \ldots, \quad N_{2(m-k)+1}(x) = x^k + \ldots$$

her. Für $k = m$ erhält man so speziell

$$\mathrm{Grad}(Z_1) \leqq m, \quad \mathrm{Grad}(N_1) = m,$$

$$Z_1(x) = \Big(\sum_{i=0}^{m} \alpha_{2i+1} \Big) x^m + \ldots, \quad N_1(x) = x^m + \ldots .$$

Die einmalige Anwendung der Gleichungen (7.4.41) für $k = 1$ liefert dann auch im vorliegenden Fall die Aussage (i).

Nun zu den Behauptungen (ii), (iii): Nach (7.4.39) hat man

$$\begin{pmatrix} P(x) \\ Q(x) \end{pmatrix} = \begin{pmatrix} Z_{0,n}(x) \\ N_{0,n}(x) \end{pmatrix}$$

und weiter nach (7.4.28), (7.4.34) für $0 \leqq i < n$

$$\begin{pmatrix} Z_{0,n}(x) \\ N_{0,n}(x) \end{pmatrix} = M_{0,n}(x) \begin{pmatrix} \alpha_n \\ 1 \end{pmatrix} = M_{0,i}(x)\, M_{i,i+1}(x)\, M_{i+1,n}(x) \begin{pmatrix} \alpha_n \\ 1 \end{pmatrix}$$

$$= M_{0,i}(x)\, M_{i,i+1}(x) \begin{pmatrix} Z_{i+1,n}(x) \\ N_{i+1,n}(x) \end{pmatrix}.$$

Wegen

$$M_{i,i+1}(x_i) = \begin{pmatrix} \alpha_i & 0 \\ 1 & 0 \end{pmatrix}$$

wird

$$M_{i,i+1}(x_i) \begin{pmatrix} Z_{i+1,n}(x_i) \\ N_{i+1,n}(x_i) \end{pmatrix} = Z_{i+1,n}(x_i) \begin{pmatrix} \alpha_i \\ 1 \end{pmatrix} = Z_{i+1}(x_i) \begin{pmatrix} \alpha_i \\ 1 \end{pmatrix}$$

und daher insgesamt

$$\text{(7.4.44)} \quad \begin{pmatrix} P(x_i) \\ Q(x_i) \end{pmatrix} = \begin{cases} \begin{pmatrix} Z_{0,n}(x_n) \\ N_{0,n}(x_n) \end{pmatrix} & \text{für } i = n, \\ Z_{i+1}(x_i) \begin{pmatrix} Z_{0,i}(x_i) \\ N_{0,i}(x_i) \end{pmatrix} & \text{für } 0 \leqq i < n. \end{cases}$$

Hieraus folgert man unter Berücksichtigung des Hilfssatzes (7.4.29), (iii)

$$Q(x_n) \neq 0, \quad R(x_n) = S_{0,n}(x_n) = f_n$$

sowie für $0 \leqq i < n$, falls $Z_{i+1}(x_i) \neq 0$ ist,

$$Q(x_i) = Z_{i+1}(x_i)\, N_{0,i}(x_i) \neq 0, \quad R(x_i) = S_{0,i}(x_i) = f_i .$$

Zu beachten bleibt, daß sich für $0 \leqq i < n$ im Fall $Z_{i+1}(x_i) = 0$ zumindest

$$P(x_i) = f_i\, Q(x_i) \quad (= 0)$$

ergibt.

Anmerken wollen wir den bereits mitbewiesenen

(7.4.45) **Zusatz.** Die Bedingungen

$$Z_{i+1}(x_i) \neq 0 \quad (i = 0, 1, \ldots, n-2)$$

sind hinreichend für die Lösbarkeit der Interpolationsaufgabe (7.4.1).

Zur numerischen Berechnung eines Wertes $R(\alpha)$ kann man sich auf die linearen Rekursionen (7.4.41) stützen; dabei fallen $2n$ Multiplikationen und eine Division an. Zweckmäßiger ist es jedoch, $R(\alpha) = S_0(\alpha)$ mittels (7.4.31), d.h. über die nicht-lineare Rekursion

$$(7.4.46) \quad \begin{cases} S_n(\alpha) = \alpha_n , \\ S_{k-1}(\alpha) = \alpha_{k-1} + \dfrac{\alpha - x_{k-1}}{S_k(\alpha)} \quad (k = n, n-1, \ldots, 1) \end{cases}$$

zu ermitteln. Dazu sind lediglich n Divisionen nötig.

Natürlich kann man (7.4.46) (bei nicht zu großem n) auch dazu benutzen, R explizit anzugeben; durch Einsetzen erhält man R in der Gestalt eines endlichen Kettenbruchs.

(7.4.47) **Beispiel.** Wir wollen die Interpolationsaufgabe (7.4.1) bezüglich $l = 2$, $m = 1$ und der vorgegebenen Werte

x_i	0	1	2	3
f_i	3	1	1	$\frac{3}{2}$

– vgl. auch Beispiel (7.4.23) – lösen. Das Schema (7.4.25) der inversen Differenzen lautet hier

x_i	$\varphi_0(x_i)$	$\varphi_1(x_0, x_i)$	$\varphi_2(x_0, x_1, x_2)$	$\varphi_3(x_0, x_1, x_2, x_3)$
0	3			
1	1	$-\frac{1}{2}$		
2	1	-1	-2	
3	$\frac{3}{2}$	-2	$-\frac{4}{3}$	$\frac{3}{2}$,

dabei stehen die α_i in der oberen Schrägzeile. Nach (7.4.46) ergibt sich $R^{2,1}(x) = S_0(x)$ als

$$R^{2,1}(x) = 3 + \cfrac{x-0}{-\frac{1}{2} + \cfrac{x-1}{-2 + \cfrac{x-2}{\frac{3}{2}}}} .$$

Ein Auflösen der Brüche – bzw. Anwendung der Rekursionen (7.4.41) – liefert

$$R^{2,1}(x) = \frac{x^2 - 2x + 3}{x + 1} .$$

Durch Einsetzen der Stützstellen überzeugt man sich von den Gleichungen

$$R^{2,1}(x_i) = f_i \quad (i = 0, 1, 2, 3) ,$$

mithin ist $R^{2,1}$ Lösung der rationalen Interpolationsaufgabe.

Wie bereits in (7.4.7) erwähnt, ist die Aufgabe (7.4.1) bezüglich $l = m = 1$ ($n = 2$) und der oben erwähnten $(x_0, f_0), (x_1, f_1), (x_2, f_2)$ nicht lösbar; dennoch sind die entsprechenden φ_k erklärt. Gemäß (7.4.46) erhalten wir $S_0(x) = R^{1,1}(x)$ in der Form

$$R^{1,1}(x) = 3 + \cfrac{x-0}{-\frac{1}{2} + \frac{x-1}{-2}} = 3 - 2\frac{x}{x} = \frac{x}{x} ;$$

die reduzierte Funktion $\hat{R}^{1,1}(x) = 1$ löst nicht (7.4.1).

In Ergänzung und Erweiterung vermerken wir zu (7.4.45) den

(7.4.48) **Zusatz.** Die Bedingungen

$$S_{i+1}(x_i) \neq 0 \quad (i = 0, 1, \ldots, n-2)$$

sind hinreichend für die Lösbarkeit der Interpolationsaufgabe (7.4.1).

Zum *Beweis* formulieren wir mehr ins Einzelne gehend den

(7.4.49) **Hilfssatz.** *Es sei* $i \in \{0, 1, \ldots, n-1\}$ *sowie* $S_{i+1}(x_i) \neq 0$. *Dann gilt für* $k = 0, 1, \ldots, i$

$$S_k(x_i) = \varphi_k(x_0, \ldots, x_{k-1}, x_i) \quad (\neq 0) ,$$

mithin insbesondere

$$S_0(x_i) = \varphi_0(x_i) = f_i ;$$

dabei lassen wir ausdrücklich zu, daß $S_k(x_i)$ *für gewisse* $k \in \{i+2, \ldots, n-1\}$ *verschwindet, also in* (7.4.31) *mit dem Wert* ∞ *gerechnet werden muß.*

Der *Beweis* verläuft analog demjenigen zu Hilfssatz (7.4.29), (iii), wo wir im dortigen Spezialfall $i = n$ eine ähnliche Aussage gemacht haben; die Durchführung sei dem Leser als Übungsaufgabe 7.10, (iii) empfohlen.

Die Verwandtschaft zwischen der Newtonschen Methode zur Ermittlung des Interpolationspolynoms und dem Kettenbruchverfahren zur Berechnung der rationalen Interpolierenden ist aus den Rekursionen (7.1.16) und (7.4.46) ersichtlich. Insbesondere tritt auch nach (7.4.46) beim letzten Schritt der numerisch günstige Fall der Fehlerdämpfung ein, sofern $|(\alpha - x_0)\, S_1(\alpha)^{-1}|$ wesentlich kleiner als $|f_0|$ ist, also sicher dann, wenn f_0 von Null verschieden ist und α hinreichend nahe bei x_0 liegt.

Um einen solchen Effekt auch für ein α, das in der Nähe einer anderen Stützstelle liegt, zu erreichen, sind wie bei der Newton-Interpolation die Stützstellen umzunumerieren. Dabei ergibt sich hier gegenüber der Polynom-Interpolation die Schwierigkeit, daß die φ_i für $i \geqq 3$ nicht symmetrisch in ihren Argumenten sind, so daß zur Anwendung der Rekursion (7.4.46) die zu der geänderten Stützstellen-Anordnung gehörenden φ_i neu berechnet werden müßten.

Um dies zu vermeiden, geht man von den inversen Differenzen (7.4.24) zu den *reziproken Differenzen* über, die man durch

$$(7.4.50)\quad \begin{cases} \rho_{2k}(x_0, x_1, \ldots, x_{2k}) = \sum_{i=0}^{k}{}' \varphi_{2i}(x_0, x_1, \ldots, x_{2i}) & \left(0 \leqq k \leqq \left[\frac{n}{2}\right]\right), \\ \rho_{2k+1}(x_0, \ldots, x_{2k+1}) = \sum_{i=0}^{k}{}' \varphi_{2i+1}(x_0, \ldots, x_{2i+1}) & \left(0 \leqq k \leqq \left[\frac{n-1}{2}\right]\right) \end{cases}$$

einführt. Wie man mit (7.4.24) leicht bestätigt, genügen diese reziproken Differenzen der Rekursion

$$(7.4.51)\quad \begin{cases} \rho_0(x_i) = f_i \quad (0 \leqq i \leqq n), \\ \rho_1(x_0, x_i) = \dfrac{x_0 - x_i}{f_0 - f_i} \quad (1 \leqq i \leqq n), \\ \rho_k(x_0, \ldots, x_{k-1}, x_i) = \rho_{k-2}(x_0, \ldots, x_{k-2}) \\ \qquad + \dfrac{x_i - x_{k-1}}{\rho_{k-1}(x_0, \ldots, x_{k-2}, x_i) - \rho_{k-1}(x_0, \ldots, x_{k-2}, x_{k-1})} \\ (2 \leqq k \leqq i \leqq n). \end{cases}$$

Man kann die reziproken Differenzen auch durch (7.4.51) und anschließend die inversen Differenzen durch

$$(7.4.52)\quad \begin{cases} \varphi_0(x_i) = \rho_0(x_i) \quad (0 \leqq i \leqq n), \\ \varphi_1(x_0, x_i) = \rho_1(x_0, x_i) \quad (1 \leqq i \leqq n), \\ \varphi_k(x_0, \ldots, x_{k-1}, x_i) = \rho_k(x_0, \ldots, x_{k-1}, x_i) - \rho_{k-2}(x_0, \ldots, x_{k-2}) \\ (2 \leqq k \leqq i \leqq n) \end{cases}$$

definieren; man zeigt leicht, daß diese derart bestimmten φ_i die Rekursion (7.4.24) erfüllen, folglich mit den früher eingeführten inversen Differenzen übereinstimmen. Dadurch erkennt man auch, daß bei vorgegebenen Argumenten die φ_i und die ρ_i immer gleichzeitig als komplexe Zahlen definiert sind oder nicht. Für die folgenden Überlegungen wollen wir etwas weitergehender als bisher voraussetzen, daß die φ_i bzw. ρ_i für sämtliche in Betracht zu ziehende Argumentfolgen wohldefiniert sind.

Wichtig im Hinblick auf das Verfahren (7.4.46) ist der

(7.4.53) **Satz.**

(i) *Die reziproken Differenzen sind symmetrisch in ihren Argumenten, d. h. für jede Permutation* $(i_0, i_1, \ldots, i_k)$ *der Zahlen* $(0, 1, \ldots, k)$ *gilt*

$$\rho_k(x_{i_0}, x_{i_1}, \ldots, x_{i_k}) = \rho_k(x_0, x_1, \ldots, x_k) .$$

(ii) *Die reziproken Differenzen genügen der Rekursion*

$$(7.4.54)\quad \begin{cases} \rho_0(x_i) = f_i \quad (0 \leqq i \leqq n) , \\ \rho_1(x_i, x_{i+1}) = \dfrac{x_i - x_{i+1}}{f_i - f_{i+1}} \quad (0 \leqq i \leqq n-1) , \\ \rho_k(x_i, \ldots, x_{i+k}) = \rho_{k-2}(x_{i+1}, \ldots, x_{i+k-1}) \\ \qquad + \dfrac{x_i - x_{i+k}}{\rho_{k-1}(x_i, \ldots, x_{i+k-1}) - \rho_{k-1}(x_{i+1}, \ldots, x_{i+k})} \\ (0 \leqq i \leqq i+k \leqq n,\ k \geqq 2) . \end{cases}$$

Beweis.

(i) Wir zeigen die Symmetrie von $\rho_{2k}(x_0, x_1, \ldots, x_{2k})$. Dazu wenden wir den Satz (7.4.40) für $n = 2k$ bezüglich der beiden Interpolationsaufgaben

$$(*)\quad (x_0, f_0), (x_1, f_1), \ldots, (x_{2k}, f_{2k}) ,$$

$$(**)\quad (x_{i_0}, f_{i_0}), (x_{i_1}, f_{i_1}), \ldots, (x_{i_{2k}}, f_{i_{2k}})$$

an. P, Q bzw. $\tilde{P}$, $\tilde{Q}$ seien die dort in (i) genannten Polynome. $\rho_{2k}(x_0, \ldots, x_{2k})$ bzw. $\rho_{2k}(x_{i_0}, \ldots, x_{i_{2k}})$ ist gemäß Konstruktion der Koeffizient der höchsten Potenz x^k von P bzw. $\tilde{P}$, 1 derjenigen der höchsten Potenz x^k von Q bzw. $\tilde{Q}$. Trivialerweise lösen sowohl P, Q als auch $\tilde{P}$, $\tilde{Q}$ das linearisierte Problem zu (*); daher ist auf Grund der Eindeutigkeitsaussage (7.4.3), (ii) für $x \in \mathbb{C}$

$$P(x)\,\tilde{Q}(x) = \tilde{P}(x)\,Q(x) .$$

Dividiert man diese Gleichung durch x^{2k} und läßt anschließend $x \to \infty$ streben, so erhält man

$$\rho_{2k}(x_0, \ldots, x_{2k}) = \rho_{2k}(x_{i_0}, \ldots, x_{i_{2k}}) .$$

Man erkennt, daß sich die Symmetrie von $\rho_{2k+1}(x_0, \ldots, x_{2k+1})$ analog beweisen läßt.

(ii) Die beiden ersten Zeilen von (7.4.54) sind nach (7.4.51) klar. Zur Begründung der dritten Zeile verwenden wir neben der Aussage (i) abermals (7.4.51); demnach wird, wie behauptet,

$$\rho_k(x_i, \ldots, x_{i+k}) = \rho_k(x_{i+1}, \ldots, x_{i+k}, x_i)$$

$$= \rho_{k-2}(x_{i+1}, \ldots, x_{i+k-1}) + \frac{x_i - x_{i+k}}{\rho_{k-1}(x_{i+1}, \ldots, x_{i+k-1}, x_i) - \rho_{k-1}(x_{i+1}, \ldots, x_{i+k})}$$

$$= \rho_{k-2}(x_{i+1}, \ldots, x_{i+k-1}) + \frac{x_i - x_{i+k}}{\rho_{k-1}(x_i, \ldots, x_{i+k-1}) - \rho_{k-1}(x_{i+1}, \ldots, x_{i+k})}$$

Wir kommen abschließend noch einmal auf das Kettenbruchverfahren (7.4.46) zurück, dem wir unter Benutzung der reziproken Differenzen die folgende endgültige Form geben:

Zunächst berechnet man die reziproken Differenzen gemäß der Vorschrift (7.4.54); hierzu verwendet man das

(7.4.55) **Schema.**

$$\begin{array}{llll}
x_0 & f_0 = \rho_0(x_0) & & \\
 & & \rho_1(x_0, x_1) & \\
x_1 & f_1 = \rho_0(x_1) & & \rho_2(x_0, x_1, x_2) \\
 & & \rho_1(x_1, x_2) & \\
x_2 & f_2 = \rho_0(x_2) & & \rho_2(x_1, x_2, x_3) \\
 & & \rho_1(x_2, x_3) & \\
x_3 & f_3 = \rho_0(x_3) & & \\
\vdots & \vdots & \vdots & \vdots
\end{array}$$

Sodann ermittelt man für ein vorgegebenes $\alpha \in \mathbb{C}$ den Wert $R(\alpha)$ mit Hilfe des

(7.4.56) *Algorithmus.*

$$\alpha_k := \begin{cases} \rho_0(x_0) & (k = 0), \\ \rho_1(x_0, x_1) & (k = 1), \\ \rho_k(x_0, \ldots, x_k) - \rho_{k-2}(x_0, \ldots, x_{k-2}) & (2 \leqq k \leqq n), \end{cases}$$

$$S_n := \alpha_n$$

$$S_{k-1} := \alpha_{k-1} - \frac{\alpha - x_{k-1}}{S_k} \quad (k = n, n-1, \ldots, 1),$$

$$R(\alpha) := S_0 ;$$

dabei stehen die für $0 \leqq k \leqq n$ benötigten $\rho_k(x_0, x_1, \ldots, x_k)$ in der obersten Schrägzeile des Schemas (7.4.55).

Der Algorithmus in dieser Form zwingt bei Umnumerierung der Stützstellen nicht zur Neuberechnung der zugehörigen reziproken Differenzen; wegen der Symmetrie der ρ_k ist ein analoges Vorgehen wie beim Newton-Algorithmus (7.1.16) möglich.

Mit Satz (7.4.40) und dem darauf basierenden Algorithmus (7.4.56) wird die Interpolationsaufgabe (7.4.1) nur für spezielle Zähler- und Nennergrade, nämlich für $l = m$ oder $l = m + 1$ gelöst. Zu diskutieren bleiben daher in diesem Zusammenhang die Fälle, daß $l > m + 1$ oder $l < m$ gefordert ist.

Wir beginnen mit dem Fall $l = m + k$, $k \geqq 2$. Hier betrachten wir mit P_0^{k-1}, der Lösung des Problems

$$(7.4.57) \quad \begin{cases} \operatorname{Grad}(P_0^{k-1}) \leqq k - 1 , \\ P_0^{k-1}(x_i) = f_i \quad (i = 0, 1, \ldots, k - 1) \end{cases}$$

die „reduzierte" rationale Interpolationsaufgabe

$$(7.4.58) \quad \tilde{R}(x_i) = \tilde{f}_i := \frac{f_i - P_0^{k-1}(x_i)}{\prod_{\kappa=0}^{k-1} (x_i - x_\kappa)} \quad (i = k, k + 1, \ldots, n) .$$

Wir setzen voraus, daß diese Aufgabe im Sinne des Satzes (7.4.40) lösbar ist. Da die Anzahl der Stützstellen in (7.4.58)

$$n - k + 1 = (m + k) + m - k + 1 = 2m + 1$$

beträgt, gilt für die Lösung $\tilde{R} = \frac{\tilde{P}}{\tilde{Q}}$

$$\operatorname{Grad}(\tilde{P}) \leqq m, \quad \operatorname{Grad}(\tilde{Q}) \leqq m .$$

Auf Grund dessen erhalten wir in

$$(7.4.59) \quad R(x) := P_0^{k-1}(x) + \prod_{\kappa=0}^{k-1} (x - x_\kappa) \cdot \tilde{R}(x)$$

eine Lösung des ursprünglich gestellten Problems: Die Beziehungen (7.4.57), (7.4.58) führen nämlich unmittelbar zu der Eigenschaft

$$R(x_i) = f_i \quad (i = 0, 1, \ldots, n) ;$$

ferner gilt die Darstellung

$$R(x) = \frac{P_0^{k-1}(x)\, \tilde{Q}(x) + \prod_{\kappa=0}^{k-1} (x - x_\kappa)\, \tilde{P}(x)}{\tilde{Q}(x)} =: \frac{P(x)}{Q(x)}$$

mit $\operatorname{Grad}(P) \leqq m + k = l$, $\operatorname{Grad}(Q) = m$.

Nun zur numerischen Auswertung: zunächst ermittelt man die Koeffizienten

$$(7.4.60) \quad b_\kappa = \Delta f[x_0, x_1, \ldots, x_\kappa] \quad (\kappa = 0, 1, \ldots, k - 1)$$

des Polynoms P_0^{k-1} in der Newtonschen Darstellung (7.1.11), indem man die dividierten Differenzen

$$\Delta f[x_i, x_{i+1}, \ldots, x_{i+j}] \quad (0 \leqq i \leqq i+j \leqq k-1)$$

in der Form des Schemas (7.1.15) berechnet. Danach erweitert man dieses Schema für $i \in \{k, k+1, \ldots, n\}$ jeweils um eine Schrägzeile, bestehend aus

$$f_i = \Delta f[x_i], \Delta f[x_{k-1}, x_i], \Delta f[x_{k-2}, x_{k-1}, x_i], \ldots, \Delta f[x_0, \ldots, x_{k-1}, x_i] .$$

Auf Grund der Newtonschen Interpolationsformel (7.1.11) folgt

$$f_i = P_0^{k-1}(x_i) + \Delta f[x_0, \ldots, x_{k-1}, x_i] \prod_{\kappa=0}^{k-1} (x_i - x_\kappa) ,$$

so daß gemäß der Definition in (7.4.58) mit $\Delta f[x_0, \ldots, x_{k-1}, x_i]$ bereits $\tilde{f}_i$ berechnet ist. Im Anschluß daran bestimmt man für vorgegebenes $\alpha \in \mathbb{C}$

$$(7.4.61) \qquad b_k := \tilde{R}(\alpha)$$

mittels des Algorithmus (7.4.56). Gemäß (7.4.59) hat man mit den b_κ aus (7.4.60), (7.4.61) die Beziehung

$$R(\alpha) = b_0 + b_1(\alpha - x_0) + b_2(\alpha - x_0)(\alpha - x_1) + \ldots + b_k(\alpha - x_0)(\alpha - x_1)\cdot\ldots\cdot(\alpha - x_{k-1})$$

demgemäß man abschließend $R(\alpha)$ mit einem Horner-ähnlichen Algorithmus – vgl. Übungsaufgabe 7.2 – ermittelt.

Es bleibt der Fall $l < m$. Hierzu sei (ohne Einschränkung)

$$f_0 = f_1 = \ldots = f_{k-1} = 0, \quad f_k, f_{k+1}, \ldots, f_n \neq 0$$

angenommen. Ist dabei $k = n+1$, so wird (7.4.1) durch $R = 0$ gelöst. Andernfalls ist (7.4.1) sicher nur dann lösbar, wenn $k \leqq l$ ist; dies sei für das weitere vorausgesetzt. Wir betrachten dann mit $\hat{l} = m$, $\hat{m} = l - k$ das Problem

$$(7.4.62) \qquad \begin{cases} \hat{R} = \dfrac{\hat{P}}{\hat{Q}} , \quad \mathrm{Grad}(\hat{P}) \leqq \hat{l}, \ \mathrm{Grad}(\hat{Q}) \leqq \hat{m}, \ \hat{Q} \neq 0 , \\ \hat{R}(x_i) = \hat{f}_i := \dfrac{1}{f_i} \cdot \prod_{\kappa=0}^{k-1} (x_i - x_\kappa) \ (\neq 0) \quad (i = k, k+1, \ldots, n) . \end{cases}$$

Dieses kann wegen $\hat{l} \geqq \hat{m} + 1$ mit den bereits besprochenen Methoden behandelt werden. Besitzt es eine Lösung $\hat{R}$, so erfüllt offenbar

$$(7.4.63) \qquad R(x) := \prod_{\kappa=0}^{k-1} (x - x_\kappa)\, \hat{R}(x)^{-1} = \frac{\prod_{\kappa=0}^{k-1} (x - x_\kappa)\, \hat{Q}(x)}{\hat{P}(x)}$$

die ursprüngliche Aufgabe (7.4.1).

Für die rationale Interpolierende einer hinreichend oft differenzierbaren Funktion gibt es eine Restglieddarstellung von ähnlicher Gestalt wie bei der Polynominterpolation, vgl. hierzu Milne-Thomson [39], S. 116. Diese Restglieddarstellung ist jedoch bereits in einfachen Fällen nur schwer für eine Fehlerabschätzung verwendbar.

Wie die Praxis gezeigt hat, ist die rationale Interpolation der Polynom-Interpolation in solchen Fällen überlegen, in denen die zu interpolierende Funktion nicht mehr ganz-holomorph, sondern nur noch meromorph ist. Diese Erfahrung bestätigt das nachstehende

(7.4.64) **Beispiel.** Wir interpolieren die Funktion

$$f(x) = \begin{cases} \dfrac{x}{\sin x} & (x \in]-\pi, \pi[\ , \neq 0) \\ 1 & (x = 0) \end{cases}$$

an den Stützstellen

$$x_0 = \frac{2\pi}{3}\ ,\ x_{i+1} = \frac{1}{2}\,x_i \quad (i = 0, 1, \ldots, 4)\ ,$$

um hiermit näherungsweise den Funktionswert $f(0)$ zu bestimmen. Da die Stelle $\hat{x} = 0$ außerhalb der Stützstellen liegt, spricht man hierbei von *Extrapolation.*

Bei Anwendung des Nevilleschen Algorithmus ergeben sich folgende Werte für $P_s^{\mu}(0)$

f_s	$P_s^1(0)$	$P_s^2(0)$	$P_s^3(0)$	$P_s^4(0)$	$P_s^5(0)$
2,418399					
	0,000000				
1,209200		1,180261			
	0,885196		0,981157		
1,047198		1,006045		1,000741	
	0,975833		0,999517		0,999982
1,011515		1,000333		1,000006	
	0,994208		0,999976		
1,002862		1,000020			
	0,998567				
1,000714					

Mit rationaler Interpolation gemäß dem Stoerschen Verfahren ergibt sich das Schema

f_s	$\hat{R}_s^{0,1}(0)$	$\hat{R}_s^{1,1}(0)$	$\hat{R}_s^{1,2}(0)$	$\hat{R}_s^{2,2}(0)$	$\hat{R}_s^{2,3}(0)$
2,418399					
	0,806133				
1,209200		0,948750			
	0,923475		1,004615		
1,047198		0,987050		0,999962	
	0,978184		1,000179		1,000000
1,011515		0,996743		0,999999	
	0,994355		1,000010		
1,002862		0,999185			
	0,998576				
1,000714					

Ein Vergleich zeigt, daß als Näherungen für $f(0)$ nur die $P_s^2(0)$ $(s = 1, 2, 3)$ besser als die entsprechenden $\hat{R}_s^{1,1}(0)$ sind; im übrigen ist die rationale Extrapolation durchweg genauer.

Die Überlegenheit der rationalen gegenüber der Polynom-Interpolation ist besonders deutlich, sofern die Stützstellen in der Nähe eines Pols der zu interpolierenden Funktion liegen.

7.5. Spline-Interpolation

Die Spline-Interpolation geht von folgendem physikalischen Modell aus: Vorgegeben sei in $[a, b] \subset \mathbb{R}$ eine Zerlegung

$$(7.5.1) \qquad \begin{cases} Z = (x_0, x_1, \ldots, x_n), \\ a = x_0 < x_1 < \ldots < x_n = b, \end{cases}$$

ferner $f_0, f_1, \ldots, f_n \in \mathbb{R}$. Ein homogener, elastischer Stab sei in den Punkten $(x_0, f_0), (x_1, f_1), \ldots, (x_n, f_n)$, und zwar mit senkrecht zum Stab wirkenden Kräften, eingespannt. Dann ist die Biegelinie des Stabes eine zweimal stetig differenzierbare Kurve $y(x)$; diese zeichnet sich vor allen anderen durch (x_0, f_0), $(x_1, f_1), \ldots, (x_n, f_n)$ verlaufenden, zweimal stetig differenzierbaren Kurven dadurch aus, daß die Biegungsenergie und damit die Gesamtkrümmung

$$\int_a^b \frac{y''(t)^2}{(1 + y'(t)^2)^3}\, dt$$

minimal wird. Wenn man davon ausgehen kann, daß $y'(t)$ klein ist, so ist y in 1. Näherung gleich der Lösung folgender Aufgabe

(7.5.2) Minimiere

$$\int_a^b y''(t)^2\, dt \qquad (y \in C_2[a, b])$$

unter den Nebenbedingungen

$$y(x_i) = f_i \qquad (i = 0, 1, \ldots, n).$$

Zu Verallgemeinerungen von (7.5.2) gelangt man, indem man die Minimierungsaufgabe bezüglich der m-ten $(m \geqq 2)$ Ableitung formuliert und eventuell zusätzlich geeignete Randbedingungen stellt. Die nächsten Überlegungen in dieser Richtung zielen auf eine Erweiterung des Funktionenraumes $C_m[a, b]$; dabei beschränken wir uns hier auf reellwertige Funktionen.

Bekanntlich heißt eine Funktion g von $[a, b]$ in $\mathbb{R}$ *absolutstetig* oder *totalstetig*, wenn zu jedem $\epsilon > 0$ ein $\delta > 0$ existiert, so daß für jede (nicht notwendig

überdeckende) Unterteilung

$$a \leqq a_1 < b_1 \leqq a_2 < b_2 \leqq \ldots \leqq a_k < b_k \leqq b$$

die Implikation

$$\sum_{i=1}^{k} (b_i - a_i) < \delta \Rightarrow \sum_{i=1}^{k} |g(b_i) - g(a_i)| < \epsilon$$

zutrifft. Es gilt hierzu – wie man beispielsweise bei Titchmarsh [54], S. 364 nachlesen kann – die

(7.5.3) **Bemerkung.** g ist in $[a, b]$ absolutstetig genau dann, wenn

$$\begin{cases} \text{(i)} & g \text{ differenzierbar fast überall in } [a, b], \\ \text{(ii)} & g' \in \mathcal{L}_1[a, b], \text{ d.h. Lebesgue-integrabel über } [a, b], \\ \text{(iii)} & g(x) = g(a) + \int_a^x g'(t)\,dt \quad (x \in [a, b]). \end{cases}$$

Demgemäß bezeichnen wir für $m \in \mathbb{N}, \geqq 2$

$$K_m[a, b] := \{f \in C_{m-1}[a, b]: f^{(m-1)} \text{ absolutstetig, } f^{(m)} \in \mathcal{L}_2[a, b]\},$$

$$K_{m,p}[a, b] := \{f \in K_m[a, b]: f^{(\kappa)}(a) = f^{(\kappa)}(b) \quad (\kappa = 0, 1, \ldots, m-1)\}.$$

Offenbar enthält $K_{m,p}[a, b]$ gerade die über $[a, b]$ hinaus periodisch fortsetzbaren Funktionen, deren periodische Fortsetzung für alle $[c, d] \supset [a, b]$ in $K_m[c, d]$ liegt.

Selbstverständlich gilt $C_m[a, b] \subset K_m[a, b]$; aus diesem Grund mag der mit Lebesguescher Integrationstheorie nicht vertraute Leser im Folgenden stets $C_m[a, b]$ statt $K_m[a, b]$ betrachten.

Für $f, g \in K_m[a, b]$ definieren wir

$$(7.5.4) \qquad \begin{cases} (f, g)_m := \int_a^b f^{(m)}(t)\, g^{(m)}(t)\,dt, \\ \|f\|_m := \sqrt{(f, f)_m}. \end{cases}$$

Hiermit ist $(\,,\,)_m$ eine reelle, symmetrische Bilinearform mit der Eigenschaft

$$(f, f)_m \geqq 0 \quad (f \in K_m[a, b]).$$

Zusätzlich notieren wir die

(7.5.5) **Bemerkung.** $\|f\|_m = 0$ gilt genau dann, wenn f Polynom vom Grad $\leqq m-1$ ist.

Beweis. Für ein Polynom f vom Grad $\leqq m-1$ gilt $f^{(m)} = 0$ und daher $\|f\|_m = 0$ trivialerweise. Ist umgekehrt $f \in K_m[a, b]$ und $\|f\|_m = 0$, so ist nach Definition

$f^{(m)}(x)^2 = 0$ fast überall und daher auch $f^{(m)}(x) = 0$ fast überall. Gemäß (7.5.3),(iii) ist folglich $f^{(m-1)}(x)$ in ganz $[a,b]$ konstant, mithin f ein Polynom vom Grad $\leqq m-1$.

Im Folgenden sei wieder Z eine Zerlegung von $[a,b]$ gemäß (7.5.1), s eine Abbildung von $[a,b]$ in $\mathbb{R}$ und $m \in \mathbb{N}, \geqq 2$.

(7.5.6) **Definition.** s heißt *Spline-Funktion vom Grad* $2m-1$ *zur Zerlegung* Z genau dann, wenn

(i) $s \in C_{2m-2}[a,b]$,

(ii) s, eingeschränkt auf $[x_i, x_{i+1}]$, ein (von i abhängiges) Polynom vom Grad $\leqq 2m-1$ ist $(i = 0, 1, \ldots, n-1)$.

Die Gesamtheit aller derartigen Funktionen s bezeichnen wir mit $S^Z_{2m-1}[a,b]$.

Da $m \geqq 2$ und daher $2m-2 \geqq m$ ist, gilt trivialerweise

$$S^Z_{2m-1}[a,b] \subset C_m[a,b] \subset K_m[a,b] .$$

Grundlegend für die weiteren Überlegungen ist der

(7.5.7) **Satz** *(Identität von Holladay). Für* $f \in K_m[a,b]$, $s \in S^Z_{2m-1}[a,b]$ *hat man die Gleichung*

$$\|f-s\|_m^2 = \|f\|_m^2 - \|s\|_m^2 - 2\sum_{\kappa=1}^{m-1}{}' (-1)^{\kappa+1}[f^{(m-\kappa)}(x) - s^{(m-\kappa)}(x)]\, s^{(m+\kappa-1)}(x)\Big|_a^b - 2\sum_{i=1}^{n}{}' (-1)^{m+1}[f(x)-s(x)]\, s^{(2m-1)}(x)\Big|_{x_{i-1}+0}^{x_i-0} \tag{7.5.8}$$

Beweis. Durch einfache Umrechnung ergibt sich

$$\|f-s\|_m^2 = \int_a^b (f^{(m)}(x) - s^{(m)}(x))^2\, dx$$

$$= \|f\|_m^2 + 2\int_a^b (f^{(m)}(x) - s^{(m)}(x))\, s^{(m)}(x)\, dx - \|s\|_m^2 .$$

Wir beachten, daß die Funktion $f^{(m-1)} - s^{(m-1)}$ in $[a,b]$ absolutstetig und daher dort nach (7.5.3), (iii) eine Stammfunktion zu $f^{(m)} - s^{(m)}$ ist, ferner, daß $s^{(m)}$

stückweise stetig differenzierbar ist; somit ist die folgende partielle Integration gerechtfertigt:

$$\int_a^b (f^{(m)}(x) - s^{(m)}(x))\, s^{(m)}(x)\, dx$$

$$= (f^{(m-1)}(x) - s^{(m-1)}(x))\, s^{(m)}(x)\Big|_a^b - \int_a^b (f^{(m-1)}(x) - s^{(m-1)}(x))\, s^{(m+1)}(x)\, dx\,.$$

Diese Beziehung formen wir durch weitere partielle Integrationen zu

$$\int_a^b (f^{(m)}(x) - s^{(m)}(x))\, s^{(m)}(x)\, dx =$$

$$= \sum_{\kappa=1}^{m-1} (-1)^{\kappa+1} (f^{(m-\kappa)}(x) - s^{(m-\kappa)}(x))\, s^{(m+\kappa-1)}(x)\Big|_a^b$$

$$+ (-1)^{m+1} \int_a^b (f'(x) - s'(x))\, s^{(2m-1)}(x)\, dx$$

um. Beachten wir abschließend, daß

$$\int_a^b (f'(x) - s'(x))\, s^{(2m-1)}(x)\, dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} (f'(x) - s'(x))\, s^{(2m-1)}(x)\, dx$$

und daß $s^{(2m-1)}(x)$ in jedem der Intervalle $]x_{i-1}, x_i[$ eine Konstante ist, so erhalten wir

$$\int_{x_{i-1}}^{x_i} (f'(x) - s'(x))\, s^{(2m-1)}(x)\, dx = (f(x) - s(x))\, s^{(2m-1)}(x)\Big|_{x_{i-1}+0}^{x_i-0}$$

und damit die Identität (7.5.8).

Im Anschluß an die Definition (7.5.6) verwenden wir die

(7.5.9) **Bezeichnungen.**

(i) $N^Z_{2m-1}[a,b] := \{s \in S^Z_{2m-1}[a,b]\colon s^{(m+\kappa)}(a) = s^{(m+\kappa)}(b) = 0 \ (\kappa = 0, \ldots, m-2)\}$

heißt Raum der *natürlichen Splines;*

(ii) $P^Z_{2m-1}[a,b] := \{s \in S^Z_{2m-1}[a,b]\colon s^{(\kappa)}(a) = s^{(\kappa)}(b) \ (\kappa = 0, \ldots, 2m-2)\}$

heißt Raum der *periodischen Splines.*

Hierzu notieren wir den

(7.5.10) **Hilfssatz.** *Ist* $f \in K_m[a,b]$, $s \in S^Z_{2m-1}[a,b]$ *mit*

$$(7.5.11) \qquad s(x_i) = f(x_i) \quad (i = 0, 1, \ldots, n)$$

und darüber hinaus eine der folgenden Bedingungen

$$(7.5.12) \quad \begin{cases} \text{(i)} & s \in N^Z_{2m-1}[a,b], \\ \text{(ii)} & s^{(\kappa)}(a) = f^{(\kappa)}(a),\ s^{(\kappa)}(b) = f^{(\kappa)}(b) \quad (\kappa = 1, \ldots, m-1), \\ \text{(iii)} & f \in K_{m,p}[a,b],\ s \in P^Z_{2m-1}[a,b] \end{cases}$$

erfüllt, so folgt die Identität

$$(7.5.13) \qquad \|f - s\|^2_m = \|f\|^2_m - \|s\|^2_m \quad (\geqq 0).$$

Beweis. Aufgrund von (7.5.11) verschwindet in (7.5.8) der Term

$$-2 \sum_{i=1}^{n}{}' (-1)^{m+1}[f(x) - s(x)]\, s^{(2m-1)}(x) \Big|_{x_{i-1}+0}^{x_i - 0},$$

außerdem ist in jedem der Fälle (7.5.12), (i), (ii), (iii) auch der Ausdruck

$$-2 \sum_{\kappa=1}^{m-1}{}' (-1)^{\kappa+1}[f^{(m-\kappa)}(x) - s^{(m-\kappa)}(x)]\, s^{(m+\kappa-1)}(x) \Big|_a^b$$

gleich Null.

Gemäß der Fallunterscheidung in (7.5.12) formulieren wir drei Typen von *Interpolationsaufgaben:*

Vorgegeben seien $f_0, f_1, \ldots, f_n$ sowie eventuell $f_a^{(1)}, \ldots, f_a^{(m-1)}$, $f_b^{(1)}, \ldots, f_b^{(m-1)} \in \mathbb{R}$. Gesucht ist eine Funktion $s \in S^Z_{2m-1}[a,b]$, die neben der Forderung

$$(7.5.14) \qquad s(x_i) = f_i \quad (i = 0, 1, \ldots, n)$$

zusätzlich einer der folgenden drei Randbedingungen

$$(7.5.15) \quad \begin{cases} \text{(i)} & s \in N^Z_{2m-1}[a,b], \\ \text{(ii)} & s^{(\kappa)}(a) = f_a^{(\kappa)},\ s^{(\kappa)}(b) = f_b^{(\kappa)} \quad (\kappa = 1, \ldots, m-1), \\ \text{(iii)} & s \in P^Z_{2m-1}[a,b] \end{cases}$$

genügt.

Hierzu notieren wir als Existenz- und Eindeutigkeitsaussage den

(7.5.16) **Satz.** *In* $S^Z_{2m-1}[a,b]$ *besitzt die Interpolationsaufgabe* (7.5.14) *mit einer der Randbedingungen* (7.5.15), (i), (ii) *oder* (iii) *jeweils genau eine Lösung, wobei*

im Fall der Bedingung (7.5.15), (i) $n \geqq m-1$ *sowie im Fall der Forderung* (7.5.15), (iii) $f_0 = f_n$ *vorausgesetzt wird.*

Beweis. Um zunächst die *Eindeutigkeit* zu zeigen, gehen wir aus von Funktionen $s, \tilde{s} \in S^Z_{2m-1}[a,b]$, die den Gleichungen

$$s(x_i) = \tilde{s}(x_i) = f_i \quad (i = 0, 1, \ldots, n)$$

und gemeinsam einer der Bedingungen (7.5.15), (i), (ii) bzw. (iii) genügen. Dann sind bezüglich s und $f := \tilde{s}$ die Voraussetzungen des Hilfssatzes (7.5.10) erfüllt. Es folgt gemäß (7.5.13)

$$\|\tilde{s} - s\|^2_m = \|\tilde{s}\|^2_m - \|s\|^2_m \geqq 0$$

und, da man s und $\tilde{s}$ vertauschen kann, ebenso

$$\|s - \tilde{s}\|^2_m = \|s\|^2_m - \|\tilde{s}\|^2_m \geqq 0\,,$$

mithin

$$\|\tilde{s} - s\|_m = 0\,.$$

Nach (7.5.5) ist $P := \tilde{s} - s$ ein Polynom vom Grad $\leqq m-1$.

Genügen $s, \tilde{s}$ beide der Bedingung (7.5.15), (i) und ist $n \geqq m-1$, so besitzt P wegen

$$P(x_i) = \tilde{s}(x_i) - s(x_i) = 0 \quad (i = 0, 1, \ldots, n)$$

mindestens m Nullstellen und ist daher das Nullpolynom.

Falls $s, \tilde{s}$ beide die Eigenschaft (7.5.15), (ii) besitzen, so gilt

$$P^{(\kappa)}(a) = \tilde{s}^{(\kappa)}(a) - s^{(\kappa)}(a) = 0 \quad (\kappa = 0, 1, \ldots, m-1)$$

und folglich nach dem Satz von Taylor wiederum $P = 0$.

Im periodischen Fall (iii) schließlich haben wir

$$P^{(\kappa)}(a) = P^{(\kappa)}(b) \quad (\kappa = 0, 1, \ldots, m-1)\,.$$

Durch Vergleich der Taylorentwicklungen von P um $x = a$ und um $x = b$ ergibt sich, daß P periodisch mit der Periode $b-a$ und folglich konstant ist. Beachtet man noch

$$P(a) = 0\,,$$

so ist auch hier die Aussage $P = 0$, d.h. $\tilde{s} = s$ klar.

Nun zur *Lösbarkeit* der genannten Interpolationsaufgaben: Durch die Zerlegung Z sind disjunkte Teilintervalle

$$I_j := [x_{j-1}, x_j[\quad (j = 1, \ldots, n-1),\ I_n := [x_{n-1}, b]$$

von $[a,b]$ gegeben; dementsprechend bezeichnen wir

$\Omega^Z_{2m-1}[a,b] := \{f\colon f$ Abbildung von $[a,b]$ in $\mathbb{R}$, $f|_{I_j}$ Polynom vom Grad $\leqq 2m-1\}$.

Offenbar ist $\Omega^{Z}_{2m-1}[a,b]$ ein Vektorraum der Dimension 2mn. Durch

$$\varphi(f) := (f^{(\kappa)}(x_j + 0) - f^{(\kappa)}(x_j - 0))_{\substack{\kappa = 0,1,\ldots,2m-2 \\ j = 1,2,\ldots,n-1}} \quad (f \in \Omega^{Z}_{2m-1}[a,b])$$

definieren wir eine lineare Abbildung φ von $\Omega^{Z}_{2m-1}[a,b]$ in $\mathbb{R}^{(2m-1)(n-1)}$. Trivialerweise ist $S^{Z}_{2m-1}[a,b]$ der Nullraum von φ, daher folgt aus der Dimensionsformel für lineare Abbildungen – vgl. Fischer [17], S. 66 ff.

(*) $\dim S^{Z}_{2m-1}[a,b] \geqq 2nm - (2m-1)(n-1) = n - 1 + 2m.$

Weiter befassen wir uns zunächst mit dem Fall (ii); wir zeigen, daß für beliebig vorgegebene $f_0, f_1, \ldots, f_n, f_a^{(1)}, \ldots, f_a^{(m-1)}, f_b^{(1)}, \ldots, f_b^{(m-1)} \in \mathbb{R}$ eine Funktion s mit den Eigenschaften (7.5.14), (7.5.15), (ii) existiert. Hierzu betrachten wir die durch

$$\psi_2(s) := (s(x_0), \ldots, s(x_n), s'(a), \ldots, s^{(m-1)}(a), s'(b), \ldots, s^{(m-1)}(b))$$

definierte lineare Abbildung ψ_2 von $S^{Z}_{2m-1}[a,b]$ in $\mathbb{R}^{n-1+2m}$. Diese ist auf Grund der oben bewiesenen Eindeutigkeitsaussage injektiv, daher ist

$$\dim S^{Z}_{2m-1}[a,b] \leqq n - 1 + 2m,$$

also wegen (*)

(7.5.17) $\dim S^{Z}_{2m-1}[a,b] = n - 1 + 2m,$

womit die behauptete Surjektivität von ψ_2 gezeigt ist.

Wir kommen zum Fall (i): nach Definition ist $N^{Z}_{2m-1}[a,b]$ der Nullraum der linearen Abbildung $\tilde{\varphi}$ von $S^{Z}_{2m-1}[a,b]$ in $\mathbb{R}^{2(m-1)}$, definiert durch

$$\tilde{\varphi}(s) := (s^{(m)}(a), s^{(m+1)}(a), \ldots, s^{(2m-2)}(a), s^{(m)}(b), \ldots, s^{(2m-2)}(b)).$$

Auf Grund der Dimensionsformel und (7.5.17) ergibt sich unmittelbar

$$\dim N^{Z}_{2m-1}[a,b] \geqq n + 1.$$

Ferner ist im Fall $n \geqq m - 1$

$$\psi_1(s) := (s(x_i))_{i=0}^{n} \quad (s \in N^{Z}_{2m-1}[a,b])$$

eine lineare und, wie oben gezeigt, auch injektive Abbildung von $N^{Z}_{2m-1}[a,b]$ in $\mathbb{R}^{n+1}$. Hieraus folgt, ähnlich wie im Fall (ii), die Gleichung

$$\dim N^{Z}_{2m-1}[a,b] = n + 1$$

und damit die Surjektivität von ψ_1.

Den Fall (iii) behandelt man analog: zunächst stellt man

$$\dim P^{Z}_{2m-1}[a,b] \geqq n$$

fest; anschließend benutzt man, daß die lineare Abbildung

$$\psi_3(s) := (s(x_i))_{i=0}^{n-1} \quad (s \in P^Z_{2m-1}[a,b])$$

injektiv ist und im hier betrachteten Fall nach Voraussetzung $f_0 = f_n$ gilt.

Ergänzend notieren wir die

(7.5.18) **Folgerung.** Für die Lösungen s_1, s_2, s_3 der Interpolationsaufgabe (7.5.14) mit den Randbedingungen (7.5.15), (i), (ii) bzw. (iii) gilt

(i) $\|s_1\|_m = \min\{\|f\|_m : f \in K_m[a,b], \quad f(x_i) = f_i \quad (i = 0, 1, \ldots, n)\}$,

(ii) $\|s_2\|_m = \min\{\|f\|_m : f \in K_m[a,b], \quad f(x_i) = f_i \quad (i = 0, 1, \ldots, n)$,
$f^{(\kappa)}(a) = f_a^{(\kappa)}, f^{(\kappa)}(b) = f_b^{(\kappa)} \quad (\kappa = 1, \ldots, m-1)\}$,

(iii) $\|s_3\|_m = \min\{\|f\|_m : f \in K_{m,p}[a,b], f(x_i) = f_i \quad (i = 0, 1, \ldots, n)\}$.

Darüber hinaus sind die s_1, s_2, s_3 durch die angegebenen Minimalitätseigenschaften eindeutig bestimmt.

Beweis. Wir bezeichnen die auf der rechten Seite von (i), (ii) und (iii) angesprochenen Teilmengen von $K_m[a,b]$ kurz mit M_1, M_2 bzw. M_3. Ist dann $f \in M_j$, so sind bezüglich f und s_j die Voraussetzungen des Hilfssatzes (7.5.10) erfüllt; es folgt aus (7.5.13)

$$\|f\|_m^2 - \|s_j\|_m^2 \geqq 0,$$

also wie behauptet,

$$\|f\|_m \geqq \|s_j\|_m .$$

Weiter nehmen wir an, es sei $\tilde{f} \in M_j$ und

$$\|\tilde{f}\|_m = \min\{\|f\|_m : f \in M_j\} = \|s_j\|_m .$$

Dann ist, wiederum nach (7.5.13),

$$\|\tilde{f} - s_j\|_m = 0$$

und folglich $\tilde{f} - s_j$ ein Polynom vom Grad $\leqq m-1$. Hieraus schließt man, wie bereits im Beweis zu Satz (7.5.16) ausgeführt, die Identität $\tilde{f} = s_j$.

Die Aussage (7.5.18), (i), speziell im Fall $m = 2$ und für $C_2[a,b]$ an Stelle von $K_2[a,b]$ formuliert, liefert s_1 offenbar als die eindeutig bestimmte Lösung des anfangs gestellten Minimierungsproblems (7.5.2).

In den folgenden Überlegungen wollen wir uns weiter auf den besonders wichtigen Fall $m = 2$ beschränken; die Funktionen $s \in S_3^Z[a,b]$ heißen *kubische* Splines. Nach den Definitionen (7.5.9) haben die Randbedingungen (7.5.15) hier die spezielle Gestalt

$$(7.5.19) \quad \begin{cases} \text{(i)} & s''(a) = s''(b) = 0, \\ \text{(ii)} & s'(a) = f_a', \; s'(b) = f_b' \quad (f_a', f_b' \in \mathbb{R}), \\ \text{(iii)} & s^{(\kappa)}(a) = s^{(\kappa)}(b) \quad (\kappa = 0, 1, 2). \end{cases}$$

Dabei muß für die Lösbarkeit von (7.5.14) im Fall (7.5.19), (iii) wiederum selbstverständlich $f_0 = f_n$ vorausgesetzt sein, während die für den Fall (i) geforderte Bedingung $n \geqq m - 1$ wegen $m = 2$ automatisch erfüllt ist.

Unser nächstes Ziel ist eine für die numerische Berechnung geeignete Darstellung der interpolierenden Splines $s \in S_3^Z[a, b]$. Dazu zeigen wir den

(7.5.20) **Hilfssatz.** *Es sei* $s \in S_3^Z[a, b]$ *mit*

$$s(x_i) = f_i \quad (i = 0, 1, \ldots, n)$$

vorgegeben. Als „Momente" bezeichnen wir die Größen

$$M_i := s''(x_i) \quad (i = 0, 1, \ldots, n),$$

ferner setzen wir zur Abkürzung

$$h_i := x_{i+1} - x_i \quad (i = 0, 1, \ldots, n-1).$$

Dann gilt für $x \in [x_i, x_{i+1}]$ *die Darstellung*

$$(7.5.21) \quad s(x) = \frac{M_{i+1}}{6h_i}(x - x_i)^3 + \frac{M_i}{6h_i}(x_{i+1} - x)^3 + \left(\frac{f_{i+1}}{h_i} - \frac{M_{i+1}}{6}h_i\right)(x - x_i) + \left(\frac{f_i}{h_i} - \frac{M_i}{6}h_i\right)(x_{i+1} - x)$$

beziehungsweise, nach Potenzen von $(x - x_i)$ *entwickelt,*

$$s(x) = f_i + \beta_i(x - x_i) + \gamma_i(x - x_i)^2 + \delta_i(x - x_i)^3$$

mit den Parametern

$$(7.5.22) \quad \begin{cases} \beta_i = \dfrac{f_{i+1} - f_i}{h_i} - \dfrac{2M_i + M_{i+1}}{6}h_i, \\ \gamma_i = \dfrac{M_i}{2}, \\ \delta_i = \dfrac{M_{i+1} - M_i}{6h_i}. \end{cases}$$

Beweis. Im Intervall $[x_i, x_{i+1}]$ ist s'' eine inhomogen lineare Funktion mit den Werten M_i, M_{i+1} in den Endpunkten und daher

$$(7.5.23) \quad s''(x) = M_i \frac{x_{i+1} - x}{h_i} + M_{i+1}\frac{x - x_i}{h_i}.$$

Folglich läßt sich s' mit einer Konstanten A_i' in der Form

$$s'(x) = \frac{M_{i+1}}{2h_i}(x - x_i)^2 - \frac{M_i}{2h_i}(x_{i+1} - x)^2 + A_i'$$

und weiter

$$s(x) = \frac{M_{i+1}}{6h_i}(x - x_i)^3 + \frac{M_i}{6h_i}(x_{i+1} - x)^3 + A_i(x - x_i) + B_i(x_{i+1} - x)$$

schreiben. Hierbei sind A_i, B_i gewissen Konstanten mit

$$A_i' = A_i - B_i .$$

Aus den Gleichungen

$$s(x_i) = f_i, \quad s(x_{i+1}) = f_{i+1}$$

ergeben sich die Werte

$$A_i = \left(\frac{f_{i+1}}{h_i} - \frac{M_{i+1}}{6} h_i \right), \quad B_i = \left(\frac{f_i}{h_i} - \frac{M_i}{6} h_i \right),$$

womit (7.5.21) bereits nachgewiesen ist.

Nun zu (7.5.22): die behauptete Gestalt der Koeffizienten von s ermittelt man durch Anwendung des Satzes von Taylor; danach hat man

$$\beta_i = s'(x_i), \quad \gamma_i = \frac{1}{2} s''(x_i), \quad \delta_i = \frac{1}{6} s'''(x_i + 0) .$$

Somit ist die Gleichung $\gamma_i = \frac{1}{2} M_i$ unmittelbar klar. Ferner erhält man durch Differentiation von (7.5.21)

$$(7.5.24) \qquad s'(x) = \frac{M_{i+1}}{2h_i}(x - x_i)^2 - \frac{M_i}{2h_i}(x_{i+1} - x)^2 + \left(\frac{f_{i+1} - f_i}{h_i} - \frac{M_{i+1} - M_i}{6} h_i \right)$$

und hieraus die angegebene Formel für β_i; ebenso ergibt sich δ_i durch Differentiation von (7.5.23).

Bezüglich der Momente M_i beweisen wir den

(7.5.25) **Satz.** *Es sei* $s \in S_3^Z[a,b]$ *die Lösung der Interpolationsaufgabe* (7.5.14) *mit den Randbedingungen* (7.5.19), (i), (ii) *oder* (iii).

Dann erfüllen die (eindeutig bestimmten) M_i *die linearen Gleichungen*

$$(7.5.26) \qquad \begin{cases} \dfrac{h_{i-1}}{6} M_{i-1} + \dfrac{h_{i-1} + h_i}{3} M_i + \dfrac{h_i}{6} M_{i+1} = \dfrac{f_{i+1} - f_i}{h_i} - \dfrac{f_i - f_{i-1}}{h_{i-1}} \\ (i = 1, 2, \ldots, n-1), \end{cases}$$

ferner zusätzlich im Fall (i)

$$(7.5.26') \qquad M_0 = 0, \quad M_n = 0 ,$$

im Fall (ii)

$$(7.5.26'') \qquad \begin{cases} \dfrac{h_0}{3} M_0 + \dfrac{h_0}{6} M_1 = \dfrac{f_1 - f_0}{h_0} - f_a' \\ \dfrac{h_{n-1}}{6} M_{n-1} + \dfrac{h_{n-1}}{3} M_n = f_b' - \dfrac{f_n - f_{n-1}}{h_{n-1}} \end{cases}$$

sowie im Fall (iii)

$$(7.5.26''') \qquad \begin{cases} M_0 = M_n \\ \dfrac{h_{n-1}}{6} M_{n-1} + \dfrac{h_{n-1} + h_0}{3} M_n + \dfrac{h_0}{6} M_1 = \dfrac{f_1 - f_n}{h_0} - \dfrac{f_n - f_{n-1}}{h_{n-1}} \end{cases}$$

und sind aus ihnen berechenbar.

Beweis. Wegen der stetigen Differenzierbarkeit von s in [a, b] haben wir insbesondere

$$(7.5.27) \qquad s'(x_i - 0) = s'(x_i + 0) \quad (i = 1, 2, \ldots, n-1) .$$

Hieraus folgen die Gleichungen (7.5.26). Um dies zu erkennen, beachten wir zunächst einmal die Beziehung $s'(x_i + 0) = \beta_i$ sowie die Darstellung (7.5.22) von β_i. Ferner ermitteln wir aus (7.5.24), notiert bezüglich $i-1$ statt i sowie für $x = x_i$, die Gleichung

$$(7.5.28) \qquad s'(x_i - 0) = \frac{M_i}{2h_{i-1}} h_{i-1}^2 + \frac{f_i - f_{i-1}}{h_{i-1}} - \frac{M_i - M_{i-1}}{6} h_{i-1}$$

$$= \frac{f_i - f_{i-1}}{h_{i-1}} + \frac{h_{i-1}}{3} M_i + \frac{h_{i-1}}{6} M_{i-1} .$$

Nun zu den Randbedingungen: der Fall (7.5.19), (i) ist wegen $M_0 = s''(a)$, $M_n = s''(b)$ unmittelbar klar. Die Randbedingungen (ii) lauten

$$s'(a) = f'_a , \quad s'(b) = f'_b .$$

Hierin setzen wir $s'(a) = \beta_0$ gemäß der Darstellung (7.5.22) ein; entsprechend benutzen wir für $s'(b) = s'(x_n - 0)$ die Beziehung (7.5.28) bezüglich $i = n$, womit die behaupteten Gleichungen in diesem Fall gezeigt sind.

Die Beziehung $M_0 = M_n$ im Fall (iii) ist wiederum unmittelbar klar. Die letzte Gleichung ergibt sich aus

$$s'(x_0 + 0) = s'(x_n - 0) ,$$

indem man $s'(x_0 + 0) = s'(a) = \beta_0$ gemäß (7.5.22) sowie (7.5.28) bezüglich $i = n$ einsetzt.

Zum Beweis der letzten Aussage formen wir die Beziehungen (7.5.26), ..., (7.5.26‴) in geeigneter Weise um. Dazu multiplizieren wir die Gleichungen (7.5.26) mit

$$\frac{6}{h_{i-1} + h_i}$$

und verfahren ähnlich mit den übrigen. Dementsprechend definieren wir für $i = 1, \ldots, n-1$

$$\begin{cases} d_i := \dfrac{6}{h_{i-1} + h_i} \left(\dfrac{f_{i+1} - f_i}{h_i} - \dfrac{f_i - f_{i-1}}{h_{i-1}} \right) , \\ \mu_i := \dfrac{h_{i-1}}{h_{i-1} + h_i} , \quad \lambda_i := \dfrac{h_i}{h_{i-1} + h_i} = 1 - \mu_i , \end{cases}$$

ferner im Fall (7.5.19), (i)

$$\begin{cases} d_0 = d_n := 0 , \\ \mu_n = \lambda_0 := 0 , \end{cases}$$

im Fall (ii)

$$\begin{cases} d_0 := \dfrac{6}{h_0}\left(\dfrac{f_1 - f_0}{h_0} - f_a'\right), \quad d_n := \dfrac{6}{h_{n-1}}\left(f_b' - \dfrac{f_n - f_{n-1}}{h_{n-1}}\right), \\ \mu_n = \lambda_0 := 1 \end{cases}$$

und schließlich im Fall (iii)

$$\begin{cases} d_n := \dfrac{6}{h_{n-1} + h_0}\left(\dfrac{f_1 - f_n}{h_0} - \dfrac{f_n - f_{n-1}}{h_{n-1}}\right), \\ \mu_n := \dfrac{h_{n-1}}{h_{n-1} + h_0}, \quad \lambda_n := \dfrac{h_0}{h_{n-1} + h_0} = 1 - \mu_n . \end{cases}$$

Mit diesen Größen bilden wir

$$d := \begin{cases} (d_i)_{i=0}^{n} & \text{in den Fällen (i), (ii)}, \\ (d_i)_{i=1}^{n} & \text{im Fall (iii)} \end{cases}$$

und weiter

(7.5.29) $$G := \begin{pmatrix} 2 & \lambda_0 & & & & \\ \mu_1 & 2 & \lambda_1 & 0 & & \\ & \ddots & \ddots & \ddots & & \\ & & 0 & \mu_{n-1} & 2 & \lambda_{n-1} \\ & & & & \mu_n & 2 \end{pmatrix}_{(n+1,\,n+1)}$$ in den Fällen (i), (ii)

bzw.

(7.5.29′) $$G := \begin{pmatrix} 2 & \lambda_1 & 0 & \cdots & 0 & \mu_1 \\ \mu_2 & 2 & \lambda_2 & \ddots & 0 & \\ 0 & \ddots & \ddots & \ddots & & \\ \vdots & & & & & 0 \\ 0 & \cdots & 0 & \mu_{n-1} & 2 & \lambda_{n-1} \\ \lambda_n & 0 & \cdots & 0 & \mu_n & 2 \end{pmatrix}_{(n,\,n)}$$ im Fall (iii).

Offenbar erfüllt dann

$$M := \begin{cases} (M_i)_{i=0}^{n} & \text{in den Fällen (i), (ii)}, \\ (M_i)_{i=1}^{n} & \text{im Fall (iii)} \end{cases}$$

das Gleichungssystem

(7.5.30) $$GM = d .$$

Zu zeigen bleibt, daß dieses Gleichungssystem eindeutig lösbar ist; hierzu notieren wir den etwas mehr aussagenden

(7.5.31) **Hilfssatz.** *G ist invertierbar, und es gilt bezüglich der ∞-Norm, definiert in* (3.3.8), *die Abschätzung*

$$\|G^{-1}\|_\infty \leqq 1 . \tag{7.5.32}$$

Beweis. Wir schreiben G in der Form

$$G = 2I + A = 2(I + \tfrac{1}{2} A) ,$$

worin auf Grund der Eigenschaften der μ_i, λ_i

$$\|A\|_\infty \leqq 1$$

gilt. Nach Satz (3.2.13) ist infolgedessen $I + \frac{1}{2} A$, mithin auch G invertierbar. Ferner ergibt sich wegen

$$G^{-1} = \tfrac{1}{2} (I + \tfrac{1}{2} A)^{-1}$$

ebenfalls gemäß Satz (3.2.13)

$$\|G^{-1}\|_\infty \leqq \frac{1}{2} \frac{1}{1 - \frac{1}{2}\|A\|_\infty} \leqq 1 .$$

Die numerische Berechnung der M_i und damit der Lösung $s \in S_3^Z[a,b]$ der Interpolationsaufgabe (7.5.14) mit einer der Randbedingungen (7.5.19), (i), (ii) oder (iii) geschieht natürlich durch Lösen des Gleichungssystems (7.5.30). In dieser Hinsicht wichtig ist der

(7.5.33) **Satz.** *G besitzt eine* LR*-Zerlegung; die Koeffizienten von* L *sind betraglich* $\leqq 1$.

Der *Beweis* wird als Übungsaufgabe 7.12 empfohlen.

Demnach läßt sich (7.5.30) mit Gauß-Elimination bei diagonaler Pivotwahl lösen, wobei dieses Vorgehen numerisch stabil ist. Da die Matrix G schwach besetzt ist, werden hierbei nur etwa

5n Multiplikationen bzw. Divisionen in den Fällen (i), (ii)

und etwa

10n Multiplikationen bzw. Divisionen im Fall (iii)

benötigt.

Angemerkt sei, daß man die Lösung des Gleichungssystems (7.5.30) ebenfalls mit den in 6.3 bzw. 6.4 angegebenen Iterationsverfahren berechnen kann; jedoch wird der Rechenaufwand dabei insgesamt wesentlich größer als bei der Behandlung mittels Gauß-Elimination. Vgl. hierzu Übungsaufgabe 7.13.

Zur Darstellung von s kann man statt der $M_i = s''(x_i)$ die Größen $m_i := s'(x_i)$ heranziehen. Die m_i berechnet man ebenfalls durch Lösen eines Gleichungssystems, das ähnliche numerische Eigenschaften wie (7.5.30) besitzt. Hiermit befaßt sich die Übungsaufgabe 7.14.

Im Folgenden beschäftigen wir uns mit dem Fall, daß eine vorgegebene Funktion f durch kubische Splines interpoliert wird; es interessiert eine Fehlerabschätzung und das Konvergenzverhalten bei wachsender Stützstellenzahl.

Zu einer Zerlegung Z von [a, b] gemäß (7.5.1) definieren wir

$$(7.5.34)\quad \begin{cases} \delta(Z) := \max\limits_{i=0}^{n-1} (x_{i+1} - x_i) \\ \sigma(Z) := \min\limits_{i=0}^{n-1} (x_{i+1} - x_i)\,. \end{cases}$$

Ferner bezeichnen wir für eine Funktion $f \in C_0[a,b]$

$$(7.5.35)\quad \omega(f,\delta) := \sup\{|f(x) - f(x')| : x, x' \in [a,b],\ |x - x'| \leqq \delta\} \quad (\delta > 0)$$

als *Stetigkeitsmodul* von f. Da f in [a, b] gleichmäßig stetig ist, gilt

$$(7.5.36)\quad \lim_{\delta \searrow 0} \omega(f,\delta) = 0$$

Wir zeigen den

(7.5.37) **Satz.** *Es sei* $f \in C_0[a,b]$, *Z eine Zerlegung von* [a, b] *mit*

$$\frac{\delta(Z)}{\sigma(Z)} \leqq \beta \quad (\geqq 1)\,.$$

Ferner sei $s \in S_3^Z[a,b]$ *Lösung der Interpolationsaufgabe*

$$s(x_i) = f_i := f(x_i) \quad (i = 0, 1, \ldots, n)$$

mit den Randbedingungen (7.5.19), (i), (ii) *oder* (iii). *Hierbei sei im Fall* (ii) $f'_a, f'_b \in \mathbb{R}$ *vorgegeben, im Fall* (iii) f(a) = f(b) *vorausgesetzt.*

Wir behaupten: mit

$$\delta := \delta(Z),\ \sigma := \sigma(Z)$$

$$K := \begin{cases} 0 \text{ in den Fällen (7.5.19), (i), (iii)}, \\ \max\{|f'_a|, |f'_b|\} \text{ im Fall (7.5.19), (ii)} \end{cases}$$

gilt die Abschätzung

$$(7.5.38)\quad \max_{x \in [a,b]} |f(x) - s(x)| \leqq \frac{4}{3\sqrt{3}} \beta^2 [\omega(f,\delta) + K \cdot \sigma] + \frac{3}{2} \omega(f,\delta)\,.$$

Beweis. Wir verwenden für s die Darstellung (7.5.21). Beachtet man hierzu die Identität

$$\begin{aligned} &\frac{f_{i+1}}{h_i}(x - x_i) + \frac{f_i}{h_i}(x_{i+1} - x) \\ &= \frac{f_{i+1} + f_i}{2} + \frac{1}{2h_i}\{2f_{i+1}(x - x_i) - 2f_i(x - x_{i+1}) - (f_{i+1} + f_i)(x_{i+1} - x_i)\} \\ &= \frac{f_{i+1} + f_i}{2} + \frac{1}{2h_i}(f_{i+1} - f_i)(2x - x_i - x_{i+1})\,, \end{aligned}$$

so ergibt sich für $x \in [x_i, x_{i+1}]$

$$(7.5.39)\quad \begin{cases} s(x) - f(x) = \dfrac{M_{i+1}}{6}(x - x_i)\left[\dfrac{(x - x_i)^2}{h_i} - h_i\right] + \dfrac{M_i}{6}(x_{i+1} - x)\left[\dfrac{(x_{i+1} - x)^2}{h_i} - h_i\right] \\ \qquad + \left[\dfrac{1}{2h_i}(f_{i+1} - f_i)(2x - x_i - x_{i+1})\right] + \left[\dfrac{f_{i+1} + f_i}{2} - f(x)\right]. \end{cases}$$

Es sind die Summanden auf der rechten Seite von (7.5.39) abzuschätzen. Zunächst hat man

$$(7.5.40)\quad \left|\frac{f_{i+1} + f_i}{2} - f(x)\right| = \frac{1}{2}\left|f(x_{i+1}) - f(x) + f(x_i) - f(x)\right|$$
$$\leqq \omega(f, h_i) \leqq \omega(f, \delta).$$

Ferner gilt

$$|2x - x_i - x_{i+1}| = 2\left|x - \frac{x_i + x_{i+1}}{2}\right| \leqq h_i$$

und daher

$$(7.5.41)\quad \left|\frac{1}{2h_i}(f_{i+1} - f_i)(2x - x_i - x_{i+1})\right| \leqq \frac{1}{2}\omega(f, h_i) \leqq \frac{1}{2}\omega(f, \delta).$$

Zur Abschätzung des ersten Summanden in (7.5.39) diskutieren wir die Funktion

$$\eta(x) := (x - x_i)\left[h_i - \frac{(x - x_i)^2}{h_i}\right].$$

Offenbar ist η im Inneren von $[x_i, x_{i+1}]$ positiv, in den Endpunkten Null. Wie man durch Nullstellenbestimmung der Ableitung verifiziert, nimmt η ihr Maximum in

$$\tilde{x} = x_i + \frac{h_i}{\sqrt{3}}$$

an. Es folgt für $x \in [x_i, x_{i+1}]$

$$\left|(x - x_i)\left[h_i - \frac{(x - x_i)^2}{h_i}\right]\right| = \eta(x) \leqq \eta(\tilde{x}) = \frac{2}{3\sqrt{3}}h_i^2.$$

Ebenso erschließt man

$$\left|(x_{i+1} - x)\left[\frac{(x_{i+1} - x)^2}{h_i} - h_i\right]\right| \leqq \frac{2}{3\sqrt{3}}h_i^2.$$

Nimmt man (7.5.40), (7.5.41) hinzu und beachtet man wiederum $h_i \leqq \delta$, so wird

$$(7.5.42)\quad |s(x) - f(x)| \leqq \frac{3}{2}\omega(f, \delta) + \frac{4}{3\sqrt{3}}\delta^2 \cdot \frac{1}{6}\max_{i=0}^{n}|M_i|.$$

Diese Abschätzung gilt, da i beliebig aus $\{0, 1, \ldots, n-1\}$ gewählt war, bereits in ganz $[a, b]$.

Es bleibt das Maximum der $|M_i|$ abzuschätzen. Laut (7.5.30) ist

$$M = G^{-1} d$$

und folglich nach (7.5.32)

$$(7.5.43) \qquad \max_{i=0}^{n} |M_i| \leqq \max_{i=0}^{n} |d_i| .$$

Hierzu erklären wir, um Fallunterscheidungen zu vermeiden, im Fall (7.5.19), (iii)

$$d_0 := d_n .$$

Weiter haben wir zunächst für $i = 1, 2, \ldots, n-1$

$$(*) \qquad |d_i| = \left| \frac{6}{h_{i-1} + h_i} \left(\frac{f_{i+1} - f_i}{h_i} - \frac{f_i - f_{i-1}}{h_{i-1}} \right) \right| \leqq \frac{6}{\sigma^2} \omega(f, \delta) .$$

Diese Ungleichung gilt im Fall (i) trivialerweise auch für $i = 0$ und $i = n$. Im Fall (ii) schätzt man

$$|d_0|, |d_n| \leqq \frac{6}{\sigma} \left(\frac{\omega(f, \delta)}{\sigma} + K \right)$$

ab. Schließlich stellt man im Fall (iii) fest, daß (*) hier auch für $i = n$ und somit natürlich auch für $i = 0$ gilt. Danach hat man in jedem der betrachteten Fälle

$$(7.5.44) \qquad \max_{i=0}^{n} |d_i| \leqq \frac{6}{\sigma} \left(\frac{\omega(f, \delta)}{\sigma} + K \right) .$$

Um den Beweis des Satzes (7.5.37) abzuschließen, bleibt (7.5.44) in (7.5.43) und dies wiederum in (7.5.42) einzusetzen und

$$\frac{\delta}{\sigma} \leqq \beta$$

abzuschätzen.

Wir notieren als unmittelbare

(7.5.45) **Folgerung.** Es sei $f \in C_0[a, b]$. Weiter bezeichne

$$\begin{cases} Z_k = (x_0^{(k)}, x_1^{(k)}, \ldots, x_{n_k}^{(k)}) \\ a = x_0^{(k)} < x_1^{(k)} < \ldots < x_{n_k}^{(k)} = b \end{cases}$$

eine Folge von Zerlegungen des Intervalls $[a, b]$, und zwar mit den Eigenschaften

$$\begin{cases} \delta_k := \delta(Z_k) \to 0 \quad (k \to \infty) , \\ \dfrac{\delta(Z_k)}{\sigma(Z_k)} \leqq \beta \quad (k \in \mathbb{N}) . \end{cases}$$

Vorgegeben seien ferner für $k \in \mathbb{N}$ $s_k \in S_3^{Z_k}[a, b]$, die den Interpolationsbedingungen

$$s_k(x_i^{(k)}) = f(x_i^{(k)}) \quad (i = 0, 1, \ldots, n_k)$$

und den Randbedingungen (i), (ii) oder (iii) genügen mögen; dabei habe man im Fall (ii) f'_a, $f'_b \in \mathbb{R}$, im Fall (iii) sei $f(a) = f(b)$ vorausgesetzt.

Wir behaupten

$$\max_{x \in [a,b]} |f(x) - s_k(x)| \to 0 \quad (k \to \infty) .$$

Beweis. Zu beachten ist die Abschätzung (7.5.38) sowie die Konvergenzaussage (7.5.36).

Die Konvergenzgeschwindigkeit ist hierbei im wesentlichen durch $\omega(f, \delta_k)$ bestimmt. Ist beispielsweise f in [a, b] differenzierbar mit

$$|f'(x)| \leqq L \quad (x \in [a,b]) ,$$

so gilt auf Grund des Mittelwertsatzes der Differentialrechnung – vgl. (6.6.13) –

$$\omega(f, \delta_k) \leqq L \cdot \delta_k$$

und daher nach (7.5.38) wegen $\beta \cdot \sigma(Z_k) \leqq \delta_k$

$$\max_{x \in [a,b]} |f(x) - s_k(x)| \leqq \left\{ \left(\frac{4}{3\sqrt{3}} \beta^2 + \frac{3}{2} \right) L + K \cdot \beta \right\} \delta_k .$$

Angemerkt sei, daß die Behauptung der Folgerung (7.5.45) auch dann noch gültig bleibt, wenn man im Fall (ii) statt fester f'_a, $f'_b \in \mathbb{R}$ von k abhängige $f_a'^{(k)}$ $f_b'^{(k)} \in \mathbb{R}$ mit

$$K_k := \max \{ |f_a'^{(k)}|, |f_b'^{(k)}| \}, \quad K_k \cdot \delta_k \to 0 \quad (k \to \infty)$$

zuläßt.

Zu einem besseren Konvergenzverhalten bezüglich δ_k und darüber hinaus zu einer Konvergenz der Ableitungen gelangt man unter der Voraussetzung $f \in C_4[a,b]$. Hierzu beweisen wir den

(7.5.46) **Satz.** *Es sei* $f \in C_4[a,b]$, *Z eine Zerlegung von* [a,b] *mit*

$$\frac{\delta(Z)}{\sigma(Z)} \leqq \beta .$$

Es erfülle $s \in S_3^Z[a,b]$ *die Interpolationsaufgabe*

$$s(x_i) = f_i := f(x_i) \quad (i = 0, 1, \ldots, n)$$

und zusätzlich entweder – entsprechend (7.5.19), (ii) –

(7.5.47) $\quad s'(a) = f'(a), \quad s'(b) = f'(b) ,$

oder es sei – entsprechend (7.5.19), (iii) –

$$(7.5.47') \quad \begin{cases} f^{(\kappa)}(a) = f^{(\kappa)}(b) & (\kappa = 0, 1, 2) , \\ s^{(\kappa)}(a) = s^{(\kappa)}(b) & (\kappa = 0, 1, 2) \end{cases}$$

vorausgesetzt.

Wir behaupten: mit

$$\delta := \delta(Z), \quad K_4 := \max_{t \in [a,b]} |f^{(4)}(t)|$$

gilt für $x \in [a,b], \neq x_i$ $(i = 0, 1, \ldots, n)$ *die Abschätzung*

$$|s'''(x) - f'''(x)| \leqq 2\beta K_4 \delta \; ; \tag{7.5.48}$$

ferner hat man für sämtliche $x \in [a,b]$ *die Abschätzungen*

$$\begin{cases} |f''(x) - s''(x)| \leqq \frac{7}{4} \beta K_4 \delta^2 , \\ |f'(x) - s'(x)| \leqq \frac{7}{4} \beta K_4 \delta^3 , \\ |f(x) - s(x)| \leqq \frac{7}{8} \beta K_4 \delta^4 . \end{cases} \tag{7.5.49}$$

Beweis. Wir beginnen mit der Abschätzung der zweiten Ableitungen in den Stützstellen. Hierzu notieren wir die Zwischenbehauptung

$$\max_{i=0}^{n} |M_i - f''(x_i)| \leqq \frac{3}{4} \delta^2 K_4 . \tag{7.5.50}$$

Zum Beweis von (7.5.50) definieren wir

$$c := \begin{cases} (f''(x_i))_{i=0}^{n} & \text{im Fall (7.5.47)}, \\ (f''(x_i))_{i=1}^{n} & \text{im Fall (7.5.47')}, \end{cases}$$

ferner

$$e = (\epsilon_i)_{i=0\,(1)}^{n} := d - Gc$$

mit G gemäß (7.5.29) bzw. (7.5.29'). Dann ist nach (7.5.30)

$$M - c = G^{-1} e$$

und folglich wegen (7.5.32)

$$\max_{i=0}^{n} |M_i - f''(x_i)| \leqq \max_{i=0\,(1)}^{n} |\epsilon_i| . \tag{7.5.51}$$

Nach Definition gilt für $i = 1, 2, \ldots, n-1$ – dabei beachte man im Fall (7.5.47') für $i = 1$ die Gleichheit $f''(x_0) = f''(x_n)$ –

$$\epsilon_i = d_i - \mu_i f''(x_{i-1}) - 2f''(x_i) - \lambda_i f''(x_{i+1})$$

und weiter, wenn man die d_i, λ_i, μ_i einsetzt,

$$\begin{aligned} (h_{i-1} + h_i)\epsilon_i = 6 \frac{f_{i+1} - f_i}{h_i} - 6 \frac{f_i - f_{i-1}}{h_{i-1}} - h_{i-1} f''(x_{i-1}) \\ - 2(h_{i-1} + h_i) f''(x_i) - h_i f''(x_{i+1}) . \end{aligned} \tag{7.5.52}$$

Gemäß der Taylor-Entwicklung von f um $x = x_i$ erhalten wir mit einem $\xi_i^+ \in [x_i, x_{i+1}]$

$$(7.5.53) \qquad f_{i+1} = f_i + h_i f'(x_i) + \frac{h_i^2}{2} f''(x_i) + \frac{h_i^3}{6} f'''(x_i) + \frac{h_i^4}{24} f^{(4)}(\xi_i^+)$$

und daher

$$(7.5.54) \qquad 6\,\frac{f_{i+1} - f_i}{h_i} = 6f'(x_i) + 3h_i f''(x_i) + h_i^2 f'''(x_i) + \frac{h_i^3}{4} f^{(4)}(\xi_i^+) .$$

Analog ergibt sich mit einem $\xi_i^- \in [x_{i-1}, x_i]$

$$-6\,\frac{f_i - f_{i-1}}{h_{i-1}} = -6f'(x_i) + 3h_{i-1} f''(x_i) - h_{i-1}^2 f'''(x_i) + \frac{h_{i-1}^3}{4} f^{(4)}(\xi_i^-) .$$

Diese letzten beiden Gleichungen setzen wir in (7.5.52) ein; wir erhalten

$$(h_{i-1} + h_i)\epsilon_i = h_i \left(f''(x_i) - f''(x_{i+1}) + h_i f'''(x_i) + \frac{h_i^2}{4} f^{(4)}(\xi_i^+) \right)$$
$$+ h_{i-1} \left(f''(x_i) - f''(x_{i-1}) - h_{i-1} f'''(x_i) + \frac{h_{i-1}^2}{4} f^{(4)}(\xi_i^-) \right) .$$

Wir wenden den Satz von Taylor nochmals an, und zwar diesmal auf die Funktion f''; demnach wird mit $\eta_i^+ \in [x_i, x_{i+1}]$, $\eta_i^- \in [x_{i-1}, x_i]$

$$(7.5.55) \qquad \begin{cases} f''(x_i) - f''(x_{i+1}) + h_i f'''(x_i) = -\frac{1}{2} h_i^2 f^{(4)}(\eta_i^+) , \\ f''(x_i) - f''(x_{i-1}) - h_{i-1} f'''(x_i) = -\frac{1}{2} h_{i-1}^2 f^{(4)}(\eta_i^-) , \end{cases}$$

mithin

$$(h_{i-1} + h_i)\epsilon_i = h_i \left(\frac{h_i^2}{4} f^{(4)}(\xi_i^+) - \frac{h_i^2}{2} f^{(4)}(\eta_i^+) \right)$$
$$+ h_{i-1} \left(\frac{h_{i-1}^2}{4} f^{(4)}(\xi_i^-) - \frac{h_{i-1}^2}{2} f^{(4)}(\eta_i^-) \right) .$$

Folglich läßt sich für $i = 1, \ldots, n-1$

$$(7.5.56) \qquad |\epsilon_i| \leqq \frac{3}{4} \frac{h_i^3 + h_{i-1}^3}{h_i + h_{i-1}} K_4 = \frac{3}{4} (h_i^2 - h_i h_{i-1} + h_{i-1}^2) K_4 \leqq \frac{3}{4} \delta^2 K_4$$

abschätzen. – Unter der Voraussetzung (7.5.47′) gilt

$$\epsilon_n = d_n - \mu_n f''(x_{n-1}) - 2f''(x_n) - \lambda_n f''(x_1) .$$

Hieraus gewinnen wir unter Benutzung der Gleichungen $f_0 = f_n$, $f''(x_0) = f''(x_n)$ die Identität

$$(h_0 + h_{n-1})\epsilon_n = 6\,\frac{f_1 - f_0}{h_0} - 6\,\frac{f_n - f_{n-1}}{h_{n-1}} - h_{n-1} f''(x_{n-1}) - 2h_{n-1} f''(x_n)$$
$$- 2h_0 f''(x_0) - h_0 f''(x_1) .$$

Zur weiteren Umformung verwenden wir, ähnlich wie in (7.5.52) die Taylorentwicklungen von f und f'' um $x = x_0$ und um $x = x_n$. Dabei ist die Gleichung $f'(x_0) = f'(x_n)$ zu beachten. Es folgt dann wie oben die Abschätzung

$$|\epsilon_n| \leqq \frac{3}{4} \frac{h_{n-1}^2 + h_0^2}{h_{n-1} + h_0} K_4 \leqq \frac{3}{4} \delta^2 K_4 \; ;$$

hiermit ist für den Fall (7.5.47') unter Berücksichtigung von (7.5.51) die Ungleichung (7.5.50) nachgewiesen.

Es bleiben $|\epsilon_0|$ und $|\epsilon_n|$ für den Fall (7.5.47) abzuschätzen. Wir haben nach Definition

$$\epsilon_0 = d_0 - 2f''(x_0) - \lambda_0 f''(x_1)$$

und daher

$$h_0 \epsilon_0 = 6 \left(\frac{f_1 - f_0}{h_0} - f'(x_0) \right) - 2h_0 f''(x_0) - h_0 f''(x_1) \; .$$

Gemäß (7.5.54) gilt mit einem $\xi_0^+ \in [x_0, x_1]$ die Gleichung

$$6 \left(\frac{f_1 - f_0}{h_0} - f'(x_0) \right) = 3h_0 f''(x_0) + h_0^2 f'''(x_0) + \frac{h_0^3}{4} f^{(4)}(\xi_0^+) \; ,$$

ferner nach (7.5.55) mit einem $\eta_0^+ \in [x_0, x_1]$

$$h_0 f''(x_0) - h_0 f''(x_1) + h_0^2 f'''(x_0) = -\frac{h_0^3}{2} f^{(4)}(\eta_0^+)$$

und daher insgesamt

$$|\epsilon_0| = |h_0^2 \left(\frac{1}{4} f^{(4)}(\xi_0^+) - \frac{1}{2} f^{(4)}(\eta_0^+) \right)| \leqq \frac{3}{4} \delta^2 K_4 \; .$$

Ebenso wird $|\epsilon_n|$ abgeschätzt. Nimmt man (7.5.56) und wiederum (7.5.51) hinzu, so ist (7.5.50) auch für diesen Fall bewiesen.

Nun zu den im Satz (7.5.46) behaupteten Ungleichungen: Durch Differentiation von (7.5.23) erhalten wir für $x \in]x_i, x_{i+1}[\quad (i = 0, \ldots, n-1)$

$$s'''(x) - f'''(x) = \frac{M_{i+1} - M_i}{h_i} - f'''(x)$$

$$= \frac{M_{i+1} - f''(x_{i+1})}{h_i} - \frac{M_i - f''(x_i)}{h_i} + \frac{f''(x_{i+1}) - f''(x) - [f''(x_i) - f''(x)]}{h_i} - f'''(x) \; .$$

Die Taylorentwicklung von f'' um den Punkt x ergibt mit $\xi_i \in [x, x_{i+1}]$, $\eta_i \in [x_i, x]$

$$f''(x_{i+1}) - f''(x) = (x_{i+1} - x) f'''(x) + \frac{(x_{i+1} - x)^2}{2} f^{(4)}(\xi_i) \; ,$$

$$f''(x_i) - f''(x) = (x_i - x) f'''(x) + \frac{(x_i - x)^2}{2} f^{(4)}(\eta_i) \; .$$

Es folgen die Gleichungen bzw. Ungleichungen

$$\left| \frac{f''(x_{i+1}) - f''(x) - [f''(x_i) - f''(x)]}{h_i} - f'''(x) \right|$$
$$= \left| \frac{(x_{i+1} - x)^2}{2h_i} f^{(4)}(\xi_i) - \frac{(x_i - x)^2}{2h_i} f^{(4)}(\eta_i) \right|$$
$$\leqq \frac{1}{2h_i} [(x_{i+1} - x)^2 + (x_i - x)^2] K_4 \leqq \frac{1}{2} h_i K_4 \leqq \frac{1}{2} \delta K_4 .$$

Benutzt man noch, daß auf Grund von (7.5.50)

$$\left| \frac{M_{i+1} - f''(x_{i+1})}{h_i} \right| + \left| \frac{M_i - f''(x_i)}{h_i} \right| \leqq \frac{1}{\sigma} \frac{3}{2} \delta^2 K_4 \leqq \beta \frac{3}{2} \delta K_4$$

gilt und daß $\beta \geqq 1$ ist, so erhält man insgesamt, wie behauptet

$$(7.5.57) \qquad |s'''(x) - f'''(x)| \leqq \frac{1}{2} \delta K_4 + \frac{3}{2} \beta \delta K_4 \leqq 2 \beta \delta K_4 .$$

Wir kommen zur Abschätzung der zweiten Ableitung: zu einem vorgegebenen $x \in [a, b]$ wählen wir uns, was nach Definition von $\delta = \delta(Z)$ immer möglich ist, ein $i \in \{0, 1, \ldots, n\}$, so daß $|x - x_i| \leqq \frac{\delta}{2}$. Dann gilt nach (7.5.50), (7.5.57)

$$|f''(x) - s''(x)| = |f''(x_i) - s''(x_i) + \int_{x_i}^{x} (f'''(t) - s'''(t)) dt|$$
$$\leqq \frac{3}{4} K_4 \delta^2 + \frac{\delta}{2} 2\beta\delta K_4$$
$$\leqq \frac{7}{4} \beta \delta^2 K_4 .$$

Ähnlich gehen wir im Fall der ersten Ableitung vor. Wegen

$$f(x_i) - s(x_i) = 0 \quad (i = 0, 1, \ldots, n)$$

existiert nach dem Satz von Rolle in jedem Intervall $]x_i, x_{i+1}[$ ein ζ_i mit

$$f'(\zeta_i) - s'(\zeta_i) = 0 .$$

Für $x \in [x_i, x_{i+1}]$ gilt offenbar $|x - \zeta_i| < \delta$ und daher

$$|f'(x) - s'(x)| = |f'(\zeta_i) - s'(\zeta_i) + \int_{\zeta_i}^{x} (f''(t) - s''(t)) dt| \leqq \frac{7}{4} \beta K_4 \delta^3 .$$

Zum Nachweis der letzten Ungleichung wählt man zu vorgegebenem $x \in [a, b]$ wiederum ein $i \in \{0, 1, \ldots, n\}$ mit $|x - x_i| \leqq \frac{\delta}{2}$. Damit erhält man, wie behauptet,

$$|f(x) - s(x)| = \left| \int_{x_i}^{x} (f'(t) - s'(t))\, dt \right| \leqq \frac{7}{8} \beta K_4 \delta^4 .$$

Die Spline-Interpolation ist der Polynom-Interpolation insbesondere bei großer Stützstellenzahl überlegen; dies zeigt auch das

(7.5.58) **Beispiel.** Wir interpolieren die Funktion

$$f(x) = \frac{1}{1 + x^2} \quad (x \in [-5,5])$$

bezüglich der Zerlegung

$$Z = (-5, -4, -3, \ldots, 3, 4, 5)$$

durch die Splinefunktion $s \in S_3^Z[-5,5]$, die den Randbedingungen (7.5.19), (i) genügt. Zum Vergleich mit P_{10}, dem entsprechenden Interpolationspolynom 10-ten Grades – vgl. Beispiel (7.1.33) – sind in der folgenden Tabelle die Fehler beider interpolierender Funktionen für einige Zwischenstellen angegeben:

x	$s(x) - f(x)$	$P_{10}(x) - f(x)$
0,5	$0{,}2053 \cdot 10^{-1}$	$0{,}4341 \cdot 10^{-1}$
1,5	$-0{,}1035 \cdot 10^{-1}$	$0{,}7235 \cdot 10^{-1}$
2,5	$0{,}2150 \cdot 10^{-2}$	$0{,}1158 \cdot 10^{0}$
3,5	$-0{,}7900 \cdot 10^{-3}$	$-0{,}3017 \cdot 10^{0}$
4,5	$-0{,}5586 \cdot 10^{-3}$	$-0{,}1881 \cdot 10^{1}$

In der Abbildung zu Beispiel (7.1.33) würde sich $s(x)$ im Rahmen der Zeichengenauigkeit kaum von $f(x)$ unterscheiden.

Übungsaufgaben zum 7. Kapitel

Aufgabe 7.1. Vorgegeben seien $x_0, x_1, \ldots, x_n \in \mathbb{C}$, paarweise verschieden, sowie $f_0, f_1, \ldots, f_n \in \mathbb{C}$. Wir bezeichnen für

$$0 \leqq \mu \leqq s \leqq n$$

mit Q_s^μ das durch die Bedingungen

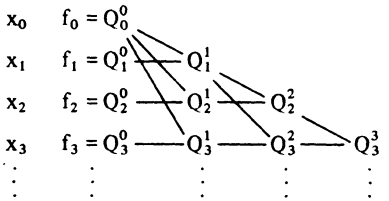
$$\begin{cases} \operatorname{Grad}(Q_s^\mu) \leqq \mu , \\ Q_s^\mu(x_i) = f_i \quad (i = 0, 1, \ldots, \mu - 1, s) \end{cases}$$

eindeutig bestimmte Polynom. Insbesondere ist $Q_n^n = P$ die Lösung der Interpolationsaufgabe (7.1.1).

Man zeige: für $x \in \mathbb{C}$ gilt

$$(*) \quad \begin{cases} Q_s^0(x) = f_s & (s = 0, 1, \ldots, n) \\ Q_s^{\mu+1}(x) = \dfrac{(x - x_\mu)\, Q_s^\mu(x) - (x - x_s)\, Q_\mu^\mu(x)}{x_s - x_\mu} \\ (s = \mu + 1, \mu + 2, \ldots, n;\ \mu = 0, 1, \ldots, n - 1) \end{cases}$$

Das Verfahren von *Aitken* benutzt die Rekursion (*), um für ein festes $x \in \mathbb{C}$ den Wert $Q_n^n(x)$ zu berechnen. Hierbei verwendet man das folgende Schema, worin zur Abkürzung $Q_s^\mu := Q_s^\mu(x)$ gesetzt sei:



Aufgabe 7.2. Vorgegeben seien $x_0, x_1, \ldots, x_{n-1}$ sowie $c_0, c_1, \ldots, c_n \in \mathbb{C}$ und hiermit

$$P(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + \ldots + c_n(x - x_0) \cdot \ldots \cdot (x - x_{n-1}),$$

ferner $\alpha \in \mathbb{C}$. Wir definieren

$$\begin{cases} s_n = c_n, \\ s_{k-1} = (\alpha - x_{k-1})\, s_k + c_{k-1} \quad (k = n, n-1, \ldots, 1). \end{cases}$$

Man zeige:

(i) $P_n(\alpha) = s_0$,

(ii) das Polynom

$$P_{n-1}(x) = s_1 + s_2(x - x_0) + \ldots + s_n(x - x_0) \cdot \ldots \cdot (x - x_{n-2})$$

hat die Eigenschaft

$$P(x) - P(\alpha) = (x - \alpha) P_{n-1}(x).$$

(iii) In Analogie zu Satz (1.4.4) formuliere man einen Algorithmus zur Entwicklung von $P(x)$ nach Potenzen von $(x - \alpha)$ bzw. zur Berechnung der $P^{(\kappa)}(\alpha)$ $(\kappa = 0, 1, \ldots, n)$.

Aufgabe 7.3. Es sei f eine k-mal differenzierbare Abbildung von $[a,b] \subset \mathbb{R}$ in $\mathbb{R}$, ferner (o.E.) $a \leqq x_0 < x_1 < \ldots < x_k \leqq b$. Bezüglich

$$(x_0, f(x_0)),\ (x_1, f(x_1)),\ \ldots,\ (x_k, f(x_k))$$

sei $\Delta f[x_0, x_1, \ldots, x_k]$ nach (7.1.10) definiert.

Man zeige, daß ein $\xi \in]x_0, x_k[$ existiert, so daß

$$\Delta f[x_0, x_1, \ldots, x_k] = \frac{1}{k!} f^{(k)}(\xi) .$$

Aufgabe 7.4 – zur quadratischen Interpolation in Funktionstafeln.

Eine 3-mal differenzierbare Abbildung f von $[a,b] \subset \mathbb{R}$ in $\mathbb{R}$ sei mit einer Schrittweite h (>0) tabelliert. Es seien

$$x_0 < x_1 < x_2 < x_3 ,\ x_i - x_{i-1} = h \quad (i = 1, 2, 3)$$

benachbarte Stützstellen der Tafel; hierzu bezeichne gemäß (7.1.6) bezüglich $f_i := f(x_i)$

$$Q(x) := \frac{1}{2} [P_0^2(x) + P_1^2(x)] .$$

Man zeige:

(i) Es gilt die Darstellung

$$Q(x) = f(x_1) + (x - x_1) \cdot \Delta f[x_1, x_2] + \frac{(x - x_1)(x - x_2)}{(x_2 - x_1)^2} \cdot \delta ,$$

mit

$$\delta := \frac{1}{2} (x_1 - x_2)^2 (\Delta f[x_0, x_1, x_2] + \Delta f[x_1, x_2, x_3]) .$$

– Diese Größen δ sind in manchen Tafeln angegeben. –

(ii) Unter der Voraussetzung

$$|f'''(t)| \leqq M \quad (t \in [a,b])$$

gilt für $x_1 < x < x_2$ die Abschätzung

$$|f(x) - Q(x)| \leqq \frac{1}{16} h^3 M .$$

Aufgabe 7.5

(i) Man verifiziere (7.2.11),

(ii) man verifiziere (7.2.12).

Aufgabe 7.6. Man beweise Satz (7.2.15).

Aufgabe 7.7. Es seien die Stützstellen

$$x_j = \frac{\pi}{2n} + j \cdot \frac{\pi}{n} \quad (j = 0, 1, \ldots, 2n - 1)$$

vorgegeben. Man zeige: die Interpolationsaufgabe

$$\begin{cases} g(x_j) = f_j \quad (j = 0, 1, \ldots, 2n-1) \\ g(x) = \dfrac{a_0}{2} + \sum\limits_{k=1}^{n-1} (a_k \cos(kx) + b_k \sin(kx)) + a_n \cos(nx) \quad (a_k, b_k \in \mathbb{C}) \end{cases}$$

ist nicht für beliebige $f_0, f_1, \ldots, f_{2n-1} \in \mathbb{C}$ lösbar; falls sie lösbar ist, so ist sie jedenfalls nicht eindeutig lösbar.

Aufgabe 7.8. Man zeige: die rationale Interpolationsaufgabe (7.4.1) bezüglich $l = 1$, $m = 2$ und

x_i	-1	0	1	2
f_i	-1	2	1	$0{,}5$

ist nicht lösbar. Sind P, Q Lösungen von (7.4.2), so hat die reduzierte Funktion

$$\hat{R} = \frac{\hat{P}}{\hat{Q}}$$

in $x_1 = 0$ einen Pol.

Aufgabe 7.9. Man berechne mit dem Stoerschen Verfahren den Wert $\hat{R}_0^{1,2}(4)$, wobei $\hat{R}_0^{1,2}$ durch die Vorschrift

x_i	0	1	3	2
f_i	$\frac{1}{3}$	1	$\frac{2}{3}$	1

bestimmt sei. – Hierbei ist gegenüber (7.4.23) die Numerierung der Stützstellen geändert; man beachte, daß nur noch eine unlösbare Aufgabe, nämlich $(I_1^{1,1})$, auftritt.

Aufgabe 7.10. Zusätzlich zu den üblichen Rechenoperationen in $\mathbb{C}$ vereinbart man in $\mathbb{C} \cup \{\infty\}$ folgende Verknüpfungen:

Für $\alpha \in \mathbb{C} \cup \{\infty\}, \neq 0$ sei

$$\frac{\alpha}{0} := \infty, \quad \alpha \cdot \infty = \infty \cdot \alpha := \infty,$$

für $\alpha \in \mathbb{C}$ sei

$$\frac{\alpha}{\infty} := 0, \quad \alpha \pm \infty = \pm \infty + \alpha := \infty;$$

hingegen sind die Ausdrücke

$$\frac{0}{0}, \quad \frac{\infty}{\infty}, \quad 0 \cdot \infty, \quad \infty \cdot 0, \quad \infty \pm \infty$$

nicht definiert. –

Vorgegeben seien $x_0, x_1, \ldots, x_n \in \mathbb{C}$, paarweise verschieden, sowie $f_0, f_1, \ldots, f_n \in \mathbb{C}$. Hierzu seien die

$$\varphi_k(x_0, x_1, \ldots, x_{k-1}, x_i) \quad (0 \leqq k \leqq i \leqq n)$$

nach (7.4.24) unter Einbeziehung der obigen Rechenoperationen, also mit Werten in $\mathbb{C} \cup \{\infty\}$, definiert; dabei gelte für $k = 0, 1, \ldots, n$

$$\alpha_k := \varphi_k(x_0, x_1, \ldots, x_k) \in \mathbb{C} \quad (\neq \infty).$$

Es seien die $Z_{k,i}(x)$, $N_{k,i}(x)$ gemäß (7.4.28) erklärt. Wir bezeichnen für $i = 0, 1, \ldots, n$

$$\tilde{D}_i := \{x \in \mathbb{C}: \forall k = 0, 1, \ldots, i \quad (N_{k,i}(x), Z_{k,i}(x)) \neq (0, 0)\}$$

und für $x \in \tilde{D}_i$

$$S_{k,i}(x) := \frac{Z_{k,i}(x)}{N_{k,i}(x)} \quad (0 \leqq k \leqq i).$$

Man zeige:

(i) Für $x \in \tilde{D}_i$ gilt – bezüglich der erweiterten Rechenoperationen –

$$S_{k-1,i}(x) = \alpha_{k-1} + \frac{x - x_{k-1}}{S_{k,i}(x)} \quad (k = i, i-1, \ldots, 1).$$

(ii) Für $i = 0, 1, \ldots, n$ hat man $x_i \in \tilde{D}_i$ und

$$S_{k,i}(x_i) = \varphi_k(x_0, \ldots, x_{k-1}, x_i) \quad (0 \leqq k \leqq i),$$

mithin insbesondere

$$S_{0,i}(x_i) = \varphi_0(x_i) = f_i.$$

Auf Grund dieser Aussage bleibt Satz (7.4.40) auch unter den hier angegebenen Voraussetzungen gültig.

(iii) Es sei $i \in \{0, 1, \ldots, n-1\}$; für $k = i+1, i+2, \ldots, n$ sei $S_k(x_i)$ im Sinne der erweiterten Operationen wohldefiniert und dabei

$$S_{i+1}(x_i) \neq 0.$$

Man zeige, daß auch für $0 \leqq k \leqq i$ die $S_k(x_i)$ definiert sind und die Gleichungen

$$S_k(x_i) = \varphi_k(x_0, \ldots, x_{k-1}, x_i) \quad (0 \leqq k \leqq i)$$

erfüllen.

Hinweis: Man zeige $S_i(x_i) = \alpha_i = S_{i,i}(x_i)$ und benutze (ii).

Aufgabe 7.11 – zum Beispiel (7.4.47).

(i) Man zeige, daß auch unter Einbeziehung der erweiterten Rechenoperationen nach Aufgabe 7.10 zu

x_i	0	1	2	3
f_i	3	1	1	$\frac{3}{2}$

nicht alle $\rho_k(x_i, \ldots, x_{i+k})$ $(0 \leqq i \leqq i+k \leqq 3)$ gemäß (7.4.54) erklärt sind.

(ii) Bezüglich der vertauschten Stützstellenanordnung

x_i	0	1	3	2
f_i	3	1	$\frac{3}{2}$	1

sind die ρ_k jedoch definiert. Man gebe die Lösung $R^{2,1}(x)$ der Interpolationsaufgabe an Hand des Algorithmus (7.4.56) an.

Aufgabe 7.12. Es sei G eine Matrix der Gestalt (7.5.29′) und hierbei

$$\lambda_i,\ \mu_i \geqq 0,\ \lambda_i + \mu_i \leqq 1 \quad (i = 1, \ldots, n).$$

Man zeige:

(i) G besitzt eine LR-Zerlegung, in $L = (l_{i,j})_{(n,n)}$ sind sämtliche

$$|l_{i,j}| \leqq 1.$$

(ii) Die Lösung des Gleichungssystems

$$GM = d$$

mit $d \in \mathbb{R}^n$ unter Benutzung der Gauß-Elimination erfordert

$10n - 14$ Multiplikationen bzw. Divisionen,

im Fall $\mu_1 = \lambda_n = 0$ nur

$5n - 4$ Multiplikationen bzw. Divisionen.

Anleitung: Man zeigt induktiv, daß nach $(k-1)$ Gauß-Eliminationsschritten $(1 \leqq k \leqq n-2)$ G in eine Matrix

$$G^{(k)} = \begin{pmatrix} r_{1,1} & \lambda_1 & 0 & \cdots & \cdots & 0 & r_{1,n} \\ 0 & \ddots & \ddots & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & \vdots \\ & 0 & r_{k,k} & \lambda_k & & & r_{k,n} \\ & & \mu_{k+1} & 2 & & & 0 \\ & & 0 & \ddots & \ddots & & \vdots \\ & & \vdots & \ddots & \ddots & \lambda_{n-1} & \vdots \\ 0 & \cdots & 0 & \omega_k & 0 \cdots 0 & \mu_n & 2 \end{pmatrix}$$

übergeht und hierin

$$r_{k,k} \geqq \frac{k+1}{k}, \quad |\omega_k| \leqq \frac{1}{k}$$

gilt.

Aufgabe 7.13. Es sei G wie in Aufgabe 7.12 vorgegeben. Man zeige:

(i) Das Gesamtschrittverfahren (6.3.4) und das Einzelschrittverfahren (6.3.6) bezüglich G konvergieren; man hat

$$\rho(A_{Ein}) \leqq \tfrac{1}{2}\,, \quad \rho(A_{Ges}) \leqq \tfrac{1}{2}\,.$$

– Es sei zusätzlich $\mu_1 = \lambda_n = 0$ angenommen. Man zeige:

(ii) Alle Eigenwerte von G sind reell,

(iii) G besitzt die Eigenschaft A – vgl. (6.4.12) –

(iv) Das Relaxationsverfahren (6.4.6) bezüglich G ist für alle $1 \leqq \omega < 2$ konvergent. Wählt man speziell

$$\omega = \omega' := \frac{4}{2+\sqrt{3}}\,,$$

so wird

$$\rho(A(\omega')) = \frac{2-\sqrt{3}}{2+\sqrt{3}} = 0{,}0718\,.$$

Anleitung zu (ii): man benutze Satz (5.4.12).

Aufgabe 7.14. Es sei $s \in S_3^Z[a,b]$ die Lösung der Interpolationsaufgabe (7.5.14) mit den Randbedingungen (7.5.19), (i), (ii) oder (iii). Es bezeichne

$$m_i := s'(x_i) \quad (i = 0, 1, \ldots, n)\,.$$

(i) Mit Hilfe von (7.2.14) zeige man für $x \in [x_i, x_{i+1}]$ die Darstellung

$$s(x) = f_i \frac{(x-x_{i+1})^2}{h_i^2}\left(1 + 2\frac{x-x_i}{h_i}\right) + m_i \frac{(x-x_i)(x-x_{i+1})^2}{h_i^2}$$
$$+ f_{i+1}\frac{(x-x_i)^2}{h_i^2}\left(1 - 2\frac{x-x_{i+1}}{h_i}\right) + m_{i+1}\frac{(x-x_i)^2(x-x_{i+1})}{h_i^2}\,.$$

(ii) Aus den Beziehungen $s''(x_i - 0) = s''(x_i + 0)$ folgere für die m_i die linearen Gleichungen

$$\left\{ h_i m_{i-1} + 2(h_{i-1}+h_i)m_i + h_{i-1}m_{i+1} = \frac{3}{h_{i-1}}h_i(f_i - f_{i-1}) + \frac{3}{h_i}h_{i-1}(f_{i+1} - f_i) \right.$$
$$\left. (i = 1, 2, \ldots, n-1) \right.$$

und ferner aus (7.5.19), (i)

$$2m_0 + m_1 = \frac{3}{h_0}(f_1 - f_0), \quad m_{n-1} + 2m_n = \frac{3}{h_{n-1}}(f_n - f_{n-1})$$

beziehungsweise aus (7.5.19), (ii)

$$m_0 = f_a'\,, \quad m_n = f_b'$$

sowie aus (7.5.19), (iii)

$$\begin{cases} m_0 = m_n \,, \\ h_{n-1} m_1 + 2(h_0 + h_{n-1}) m_n + h_0 m_{n-1} = \dfrac{3}{h_{n-1}} h_0 (f_n - f_{n-1}) + \dfrac{3}{h_0} h_{n-1} (f_1 - f_n) \end{cases}$$

Aufgabe 7.15. Es sei eine Zerlegung Z von $[a,b]$ mit

$$\frac{\delta(Z)}{\sigma(Z)} \leqq \beta$$

– vgl. (7.5.34) –, ferner $f_0, f_1, \ldots, f_n$ sowie $\tilde{f}_0, \tilde{f}_1, \ldots, \tilde{f}_n \in \mathbb{R}$ vorgegeben. Hierzu bezeichne $s, \tilde{s} \in S_3^Z[a,b]$ die durch

$$\left.\begin{aligned} s(x_i) &= f_i \\ \tilde{s}(x_i) &= \tilde{f}_i \end{aligned}\right\} \quad (i = 0, 1, \ldots, n)$$

und gemeinsam eine der Randbedingungen (7.5.19), (i), (ii) oder (iii) definierten Splines. Dabei seien im Fall (ii) für $s, \tilde{s}$ gemeinsame $f'_a, f'_b \in \mathbb{R}$ vorgegeben, im Fall (iii) $f_0 = f_n$, $\tilde{f}_0 = \tilde{f}_n$ vorausgesetzt.

Man zeige

(i) Es ist $\tilde{s} - s \in S_3^Z[a,b]$ Lösung der Interpolationsaufgabe

$$(\tilde{s} - s)(x_i) = \tilde{f}_i - f_i \quad (i = 0, 1, \ldots, n)$$

mit den Randbedingungen (7.5.19), (i), (ii) bzw. (iii), und zwar im Fall (ii) bezüglich $f'_a = f'_b = 0$.

(ii) Setzt man

$$\epsilon := \max_{i=1}^{n} |\tilde{f}_i - f_i| \,,$$

so gilt für $x \in [a,b]$ die Fehlerabschätzung

$$|\tilde{s}(x) - s(x)| \leqq \epsilon (1 + 4\beta^2) \,.$$

Anleitung zu (ii): man benutzt (7.5.21) und (7.5.43).

Literatur

[1] *J. Ahlberg, E. Nilson, J. Walsh:* The theory of splines and their applications. New York–London, Academic Press 1967.

[2] *St. Banach:* Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. Fundamenta mathematica **3** (1922).

[3] *F. L. Bauer, C. T. Fike:* Norms and exclusions theorems. Num. Math. **2** (1960), S. 137–141.

[4] *F. L. Bauer, J. Heinhold, K. Samelson, R. Sauer:* Moderne Rechenanlagen. Leitfäden der angewandten Mathematik, Bd. 5. Stuttgart, Teubner 1964.

[5] *F. L. Bauer, J. Stoer, C. Witzgall:* Absolute and monotonic norms. Num. Math. **3** (1961), S. 257–264.

[6] *I. S. Berensin, N. P. Shidkow:* Numerische Methoden 1, 2. Berlin, VEB Deutscher Verlag der Wissenschaften 1970, 1971.

[7] *K. Böhmer:* Spline-Funktionen. Teubner Studienbücher. Stuttgart, Teubner 1974.

[8] *L. Collatz:* Funktionalanalysis und numerische Mathematik. Die Grundlehren der mathematischen Wissenschaften. Berlin–Heidelberg–New York, Springer 1964.

[9] *B. Döring:* A higher order iterative method for the solution of nonlinear operator equations. In: A. Ghizzetti (Ed.): Theory and Applications of Monotone Operators. Proc. of a NATO Advanced study Institute held in Venice, Italy, 1968. Ghubbio 1968, S. 291–298.

[10] *B. Döring:* Über das Newtonsche Näherungsverfahren. Mathem. Phys. Sem. Ber. **16** (1969), S. 27–40.

[11] *B. Döring:* Einige Sätze über das Verfahren der tangierenden Hyperbeln in Banach-Räumen. Československá Akademie Věd, Aplikace Matematiky, **15** (1970), S. 418–464.

[12] *B. Döring:* Das Tschebyscheff-Verfahren in Banach-Räumen. Num. Math. **15** (1970), S. 175–195.

[13] *H. Ehrmann:* Konstruktion und Durchführung von Iterationsverfahren höherer Ordnung. Arch. Rat. Mech. Anal. 4 (1959/60), S. 65–88.

[14] *F. Erwe:* Differential- und Integralrechnung 1. Hochschultaschenbücher. Mannheim, Bibliographisches Inst. 1962.

[15] *D. K. Faddejew, W. N. Faddejewa:* Numerische Methoden der linearen Algebra. München–Wien, Oldenbourg 1970.

[16] *G. Feigl, H. Rohrbach:* Einführung in die höhere Mathematik. Berlin–Göttingen–Heidelberg, Springer 1953.

[17] *G. Fischer:* Lineare Algebra. Reinbek, rororo vieweg 17, 1975.

[18] *O. Forster:* Analysis 1, 2. Reinbek, rororo vieweg 24/31, 1975/77.

[19] *G. E. Forsythe, P. Henrici:* The cyclic Jacobi method for computing the principal values of a complex matrix. Trans. Amer. mat. Soc. **94** (1960), S. 1–23.

[20] *J. G. F. Francis:* The QR transformation, Parts I and II. Computer J. **4** (1961/62), S. 265–271, S. 332–345.

[21] *W. Greub:* Lineare Algebra. Heidelberger Taschenbücher. Berlin–Heidelberg–New York, Springer 1976.

[22] *G. Hämmerlin:* Numerische Mathematik I. Hochschultaschenbücher. Mannheim, Bibliographisches Inst. 1970.

[23] *R. W. Hamming:* Numerical Methods for Scientists and Engeneers. New York–Toronto–London, McGraw-Hill 1962.

[24] *P. Henrici:* Elemente der Numerischen Analysis I, II. Hochschultaschenbücher. Mannheim, Bibliographisches Inst. 1972.

[25] *F. B. Hildebrand:* Introduction to Numerical Analysis. New York–Toronto–London, McGraw-Hill 1956.

[26] *F. Hirzebruch, W. Scharlau:* Einführung in die Funktionalanalysis. Hochschultaschenbücher. Mannheim, Bibliographisches Inst. 1971.

[27] *A. S. Householder:* Principles of Numerical Analysis. New York–Toronto–London, McGraw-Hill 1953.

[28] *A. S. Householder:* The Theory of Matrices in Numerical Analysis. New York, Blaisdell Publishing Company 1965.

[29] *A. S. Householder:* The Numerical Treatment of a Single Nonlinear Equation. New York–Toronto–London, McGraw-Hill 1970.

[30] *M. A. Hyman:* Eigenvalues and eigenvectors of general matrices. Twelth National meeting A.C.M. Houston, Texas (1957).

[31] *E. Isaacson, H. B. Keller:* Analysis of Numerical Methods. New York, Wiley 1966.

[32] *M. A. Jenkins, J. F. Traub:* A three-stage algorithm for real polynomials using quadratic iteration. Siam J. Numer.Anal. 7 (1970), S. 545–566.

[33] *M. A. Jenkins, J. F. Traub:* A three-stage variable-shift iteration for polynomial zeros and its relation to generalized Rayleigh iteration. Num. Math. **14** (1970), S. 252–263.

[34] *F. John:* Lectures on Advanced Numerical Analysis. New York–London–Paris, Gordon and Breach 1967.

[35] *L. V. Kantorovich, G. P. Akilov:* Functional Analysis in Normed Linear Spaces. Pergamon Press, 1964.

[36] *T. Kato:* Perturbation Theory for Linear Operators. Die Grundlehren der mathematischen Wissenschaften. New York, Springer 1966.

[37] *C. Lanczos:* Applied Analysis. Englewood Cliffs, Prentice Hall 1956.

[38] *W. E. Milne:* Numerical Calculus. Princeton University Press 1949.

[39] *L. M. Milne-Thomson:* The calculus of finite differences. London, Macmillan 1965.

[40] *R. Mennicken, E. Wagenführer:* Numerische Mathematik 1. Reinbek, rororo vieweg 1977.

[41] *I. P. Natanson:* Konstruktive Funktionentheorie 1, 2, 3. Berlin, Akademie-Verlag 1955.

[42] *K. Nickel:* Die numerische Berechnung der Wurzeln eines Polynoms. Num. Math. **9** (1966), S. 80–98.

[43] *B. Noble:* Applied Linear Algebra. Englewood Cliffs, Prentice Hall 1969.

[44] *A. M. Ostrowski:* Über die Stetigkeit von charakteristischen Wurzeln in Abhängigkeit von den Matrixelementen. Jber. DMV **60** (1957), Abt. 1, S. 40–42.

[45] *A. Ralston:* A First Course in Numerical Analysis. New York, McGraw-Hill 1965.

[46] *A. Ralston, H. S. Wilf:* Mathematische Methoden für Digitalrechner I, II. München–Wien, Oldenbourg 1967, 1969.

[47] *H. Rutishauser:* Solution of eigenvalue problems with the LR-transformation. Appl. Math. Ser. nat. Bur. Stand. **49** (1958), S. 47–81.

[48] *A. Schönhage:* Zur Konvergenz des Jacobi-Verfahrens. Num. Math. **3** (1961), S. 374–380.

[49] *H. R. Schwarz, H. Rutishauser, E. Stiefel:* Numerik symmetrischer Matrizen. Leitfäden der angewandten Mathematik, Bd. 11. Stuttgart, Teubner 1968.

[50] *E. Stiefel:* Einführung in die Numerische Mathematik. Leitfäden der angewandten Mathematik, Bd. 1. Stuttgart, Teubner 1963.

[51] *J. Stoer:* Einführung in die Numerische Mathematik I, 2. Aufl. Heidelberger Taschenbücher. Berlin–Heidelberg–New York, Springer 1976.

[52] *J. Stoer, R. Bulirsch:* Einführung in die Numerische Mathematik II. Heidelberger Taschenbücher. Berlin–Heidelberg–New York, Springer 1973.

[53] *F. Stummel, K. Hainer:* Praktische Mathematik. Teubner Studienbücher. Stuttgart, Teubner 1971.

[54] *E. C. Titchmarsh:* The Theory of Functions, 2nd ed. Oxford University Press 1939.

[55] *J. F. Traub:* A class of globally convergent iteration functions for the solution of polynomial equations. Math. Comp. **20** (1966), S. 113–138.

[56] *W. Walter:* Einführung in die Potentialtheorie. Hochschulskripten. Mannheim, Bibliographisches Inst. 1971.

[57] *H. Werner:* Praktische Mathematik I, Methoden der linearen Algebra. Hochschultext. Berlin–Heidelberg–New York, Springer 1970.

[58] *H. Werner, R. Schaback:* Praktische Mathematik II, Methoden der Analysis. Hochschultext. Berlin–Heidelberg–New York, Springer 1972.

[59] *H. Wielandt:* Das Iterationsverfahren bei nicht selbstadjungierten linearen Eigenwertaufgaben. Bericht der aerodynamischen Versuchsanstalt Göttingen 44/J/37 (1944).

[60] *J. H. Wilkinson:* Rundungsfehler. Heidelberger Taschenbücher. Berlin–Heidelberg–New York, Springer 1969.

[61] *J. H. Wilkinson:* The Algebraic Eigenvalue Problem. Oxford, Clarendon Press 1965.

[62] *J. H. Wilkinson:* Note on the quadratic convergence of the cyclic Jacobi process. Num. Math. **4** (1962), S. 296–300.

[63] *J. H. Wilkinson:* Gobal convergence of tridiagonal QR-algorithm with origin shifts. Linear Algebra and its Appl. **1** (1968), S. 409–420.

[64] *J. H. Wilkinson, Ch. Reinsch:* Linear Algebra. Handbook for Automatic Computation, Vol. II. Die Grundlehren der mathematischen Wissenschaften. Berlin–Heidelberg–New York, Springer 1971.

[65] *D. M. Young:* Iterative Solutions of Large Linear Systems. New York–London, Academic Press 1971.

Kurzbiographie der Autoren

Reinhard Mennicken wurde 1935 in Köln geboren. Studium der Mathematik und Physik 1957–1962 in Köln; dort 1963 Promotion und 1969 Habilitation. 1969–1971 Dozent an der Universität Konstanz. 1971–74 Abteilungsvorsteher und Professor an der Universität Regensburg. 1972–74 Lehrstuhlvertretungen in Erlangen und Braunschweig. 1974/75 o. Professor für Mathematik an der TU Braunschweig. Seit 1975 Professor an der Universität Regensburg und Leiter einer Abteilung für Angewandte Mathematik. Forschungsschwerpunkte sind Differentialgleichungen, Funktionalanalysis und Numerische Mathematik.

Ekkehard Wagenführer wurde 1944 in Apolda/Thüringen geboren. Studium der Mathematik und Physik 1963–69 in Köln; dort 1971 Promotion. Seit 1972 Akademischer Rat beim Fachbereich Mathematik der Universität Regensburg. Forschungsschwerpunkte sind Differentialgleichungen und Numerische Mathematik.

Sachregister